On Business Analytics: Dynamic Network Analysis for Descriptive Analytics and Multicriteria Decision Analysis for Prescriptive Analytics

by

Márcia Daniela Barbosa de Oliveira

Ph.D. Dissertation in Business and Management Studies

Supervised by

Professor Doutor João Manuel Portela da Gama

Professora Doutora Dalila Benedita Machado Martins Fontes

Faculdade de Economia

Universidade do Porto

September 2015

Dedicada à minha mãe Alda
minha inesgotável fonte de inspiração, motivação e força.

Biography Note

Márcia Oliveira was born in December 3, 1986, in a small coastal city in Portugal best known for its Venice-like canals. Her curiosity and love for learning are reflected in her academic journey. During high school she pursued economics, spent a year taking art classes, and won several awards and recognitions for her academic achievements. In 2008, after finishing her undergraduate studies in management at the University of Aveiro and being granted with a merit award, she decided to explore the fascinating world of data analysis, data mining and machine learning in her Master studies at the School of Economics and Management (FEP) of the University of Porto. Since the very first class, she fell in love with the field and became passionate about finding treasures and gems in seas of data. The opportunity to immerse herself in data analysis and conduct research in this field soon appeared in 2010, when she was invited to join the Laboratory of Artificial Intelligence and Decision Support (LIAAD/INESC TEC) as a research assistant. During these five years at LIAAD/INESC TEC, Márcia developed a taste for the analysis of data and models that evolve over time, a research interest that prevails until today, and published several scientific articles on the topic. In LIAAD/INESC TEC, she was also fortunate to meet and learn from some of the brightest and most knowledgeable scientists in the fields of machine learning, statistics, optimisation and decision analysis. Their work encouraged her to continue her quest for knowledge and in September 2011 she started her Ph.D. in Business and Management Studies at FEP, under the supervision of Prof. João Gama and Prof. Dalila Fontes. Since then, she has been doing research in social network analysis, dynamic community analysis, visualisation of dynamic networks and multicriteria decision analysis. During her doctoral journey, she has also been involved in related academic tasks and roles. She collaborated in the organisation of several international scientific conferences (as publicity chair, social media chair and member of the local organisation team), attended and presented her work at scientific conferences, co-supervised Master students, gave invited lectures and seminars, acted as referee of several scientific conferences and journals, and collaborated with other scientists. These activities enabled her to grow as a researcher, and to meet, work and be surrounded with amazing and talented people all of these years.

Financial Support

My work and research were financially supported by *Fundação para a Ciência e a Tecnologia* (FCT), the Portuguese funding agency that supports science, technology, and innovation, through the Ph.D. grant SFRH/BD/81339/2011.

My research was partially funded by project "NORTE-07-0124-FEDER-000057", funded by the North Portugal Regional Operational Programme (ON.2 - O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF); by ERDF (FEDER) through the COMPETE Programme, and by national funds, through FCT, within the scope of project FCOMP-01-0124-FEDER-037281 and project KDUS (PTDC/EIA-EIA/098355/2008). I would also like to acknowledge the support of the European Commission through the project MAESTRA (Grant Number ICT-750 2013-612944).

Acknowledgements

My four-year Ph.D. journey was a rollercoaster of emotions and I went through several stages, from blissful excitement to dismay. Despite its ups and downs, this life-changing journey was worth it and deeply rewarding. I was fortunate to share it with many beautiful people, to whom I would like to give my word of appreciation and gratitude.

First of all, I would like to express my heartfelt gratitude to my supervisors, Prof. João Gama and Prof. Dalila Fontes, for always believing in me and for being a source of motivation, guidance, knowledge and support during these years. They are both brilliant researchers and I was fortunate to have had the opportunity to learn from and work with them. I am thankful to Prof. João Gama, my scientific father, for encouraging me to pursue a Ph.D. and delve into the wonderful field of social network analysis. He introduced me to the academic world and taught me, with all his patience and serenity, how to think like a researcher. He taught me how to simplify problems and focus on the big picture, how to write good scientific articles and how to calm down in stressful moments. I have learned deeply with Prof. João Gama through his words and example. He played a crucial role in my development as a researcher and I am truly grateful for all the opportunities he gave me to grow and learn. I am thankful to Prof. Dalila Fontes, my scientific mother, for introducing me to the field of multicriteria decision analysis, and for helping me improve my scientific writing and sharpen my critical reasoning. Her guidance, motivation and support were critical during these years, and I have learned so much with our discussions and conversations. I am especially grateful for her countless readings, helpful comments, wise words, optimistic attitude, and precious advice. Her teachings will follow me into my life.

A special "thank you" goes to Prof. Pavel Brazdil, the founder of LIAAD, for teaching me so much about machine learning, for being such a kind, compassionate, and broad-minded scientist, and for inspiring me in so many ways. I have learned deeply through his example. I would also like to extend my word of gratitude to Prof. Alípio Jorge, Prof. Carlos Soares, Prof. Paula Brito, Prof. Pedro Campos, Prof. Pedro Quelhas Brito, Prof. João Moreira and Prof. Luís Torgo, for sharing their knowledge with me, for their willingness and availability to help, for the opportunities to collaborate, and for their advice on research in general.

I am also indebted to Prof. Carlos Cabral Cardoso, director of the doctoral programme in

Business and Management Studies, for his help, encouragement and assistance since the very first year of my Ph.D. journey. The kind words and support of Prof. Jorge Valente were also a source of strength during the toughest years of my doctoral studies. I also owe a word of gratitude to Prof. Rui Alves for the interesting discussions and teachings: I won't ever forget the difference between "datum" and "data" thanks to him.

I was fortunate to have Cláudia Huber, David Cardoso, Diana Falcão, Nuno Rebelo, and Sam Heshmati, as classmates. Together, we've shared many good moments and smiles, and helped each other during the first year of our Ph.D. studies. They were responsible for turning my first Ph.D. year into a memorable one, and I am grateful for that. A special thank you to Nuno Rebelo for the insightful discussions about Business Analytics.

The case studies presented in chapters 3 and 5 would not have been possible without the help and collaboration of Prof. Teresa Pereira and Américo Guerreiro. Their research input, experience, business knowledge, helpful comments and discussions were essential for conducting these case studies, and I am deeply grateful to them. My word of gratitude and appreciation is also extended to the companies used as case studies and its collaborators, for providing the data and the necessary information. In particular, I would like to acknowledge the valuable help of TOYOTA Caetano de Portugal, and of Eng. Carlos Rodrigues, in providing the research scenario and the necessary conditions to conduct the study on multicriteria decision analysis.

I am also grateful to the anonymous referees for their valuable comments and suggestions on earlier versions of the published articles. Their input greatly improved the research presented here.

The friendship of Cláudio Sá and Zaigham Siddiqui made my Ph.D. journey very rewarding and I am truly grateful for the long conversations, entertaining discussions, enjoyable moments and comforting words. Thank you both for being such amazing friends and for being there when I most needed.

My gratitude also goes to Fábio Pinto, who enthusiastically revised several chapters of this dissertation and offered me helpful feedback, and to Rui Sarmento, who helped me in the experiments of Chapter 3 by running the script of the GED method in SQL. A special thank you also goes to my brother Manuel Oliveira for taking the time to read several sections of this dissertation and for offering me insightful advice on how to improve them.

I feel lucky to have had the opportunity to meet and share my working days with wonderful people at LIAAD/INESC TEC. Thank you Luís Matias, Raquel Sebastião, Pedro Almeida, Petr Kosina, Carlos Ferreira, Hadi Fanaee, Ezilda Almeida, Odair Tavares, Melissa Rodrigues, Pedro Pereira Rodrigues, Hanen Borchani, Iñigo Medialdua, Fernando Correa, Elaine Ribeiro, Ludmila Navratilova, Pedro Abreu, Fábio Pinto, Douglas Cardoso, Conceição Rocha, Vânia Almeida, João Duarte, João Vinagre, Rafael Nunes, Vítor Cerqueira, Rui Sarmento, Sónia Teixeira, Tiago Cunha, Nuno Moniz, Renato Fernandes, Abril Uriarte, Shazia Tabassum, Elena Ikonomovska, Vinícius Souza, Yusuke Sakamoto, Pawel Matuszyk, among others, for the very inspiring working

atmosphere, lively discussions about everything under the sun, amazing dinners, challenging quiz nights, scientific collaborations, and words of encouragement. You have been great labmates and colleagues, and I will miss you.

My life without friends would not be as meaningful. A special thank you goes to my friends (you know who you are), for their smiles, patience, understanding, care, emotional support, fun weekends and very long chats. You have always been there for me no matter what and you have enriched my life with true friendship. You hold a very special place in my heart and I am immensely grateful to you.

I would also like to express my warm and loving "thank you" to Igor Gama. Thank you for designing and improving the quality of some figures of this thesis. But, most importantly, thank you for brightening my days with your sweet words and affection, for being a daily source of strength, for always believing in my abilities, for encouraging me to pursue my dreams and for putting your best efforts into making these dreams come true. Your determination, strong-willed nature, confidence, and passion for life have been my daily doses of inspiration.

Last but by no means least, my warmest and deepest gratitude goes to my beloved mother, father and brother. No possible combination of words in the dictionary would be enough to faithfully express what you represent to me. You are my angels, my safe haven, my everything. Thank you for your unconditional love, sweetest smiles, comforting hugs, endless care, constant support and encouragement. Thank you for always believing that I could achieve my goals, even when they seemed impossible to me. Thank you for investing and firmly supporting my education. Thank you for your example and the values you have instilled in me. I only wish one day I can be as wise, kind-hearted and strong as you are.

Resumo

Frequentemente, os gestores têm de analisar e procurar informação relevante em enormes quantidades de dados que sustente e os oriente na tomada de decisão. Porém, encontrar a informação certa em tempo útil torna-se cada vez mais difícil à medida que a dimensão dos dados cresce, o que prejudica a capacidade dos gestores em agir e decidir rapidamente. A análise de negócios (do inglês *Business Analytics*) é uma abordagem promissora para enfrentar a sobrecarga de informação com que os gestores se deparam na atual era da informação. A análise de negócios extrai valor contido nos dados por via da descoberta de padrões e relações não triviais em grandes quantidades de dados, com recurso a técnicas quantitativas e computacionais. O objetivo é fornecer aos gestores *insights* valiosos sobre o negócio que os ajudem a tomar decisões eficazes, tendo por base a análise de dados e factos.

A análise de negócios engloba três grandes perspetivas: a descritiva, a preditiva e a prescritiva. Esta tese contribui para as perspetivas descritiva e prescritiva, uma vez que estas têm sido pouca exploradas na literatura em comparação com a perspetiva preditiva. O nosso objetivo de investigação tem, assim, uma vertente dupla: (i) por um lado, pretendemos desenvolver novas metodologias para análise descritiva, que sejam suportadas na análise de dados e capazes de extrair o valor contido em dados históricos, temporais e relacionais recolhidos pelas organizações; (ii) por outro lado, temos como intuito investigar o potencial da aplicação de abordagens híbridas de análise de decisão multicritério na resolução de problemas industriais complexos, para efeitos de análise prescritiva.

Nesta tese, contribuímos para a perspetiva descritiva da análise de negócios através do desenvolvimento de duas metodologias complementares que encontram padrões em dados temporais e relacionais, nomeadamente, (i) uma metodologia para monitorizar comunidades dinâmicas em redes que evoluem ao longo do tempo, e cuja evolução pode ser estudada a diferentes níveis de resolução temporal, e (ii) uma metodologia para visualizar a evolução deste tipo de redes ao nível dos nós ou ao nível das comunidades. A primeira metodologia explora medidas de análise de redes sociais, métodos de deteção de comunidades, e métodos para análise da evolução de comunidades, com o intuito de descobrir perfis temporais de clientes em redes dinâmicas de clientes e detetar mudanças na respetiva evolução (por exemplo, cisões e fusões de comunidades de clientes). A utilidade da metodologia proposta é demonstrada através da respetiva aplicação numa rede dinâmica de

clientes, gerada a partir dos dados disponibilizados por uma grande empresa Portuguesa. A segunda metodologia representa uma nova abordagem para a visualização de redes dinâmicas que se baseia na projeção de trajetórias espácio-temporais de nós, ou comunidades, num espaço bidimensional representativo e interpretável. Para gerar as visualizações, utilizamos tensores de terceira ordem e modelos de decomposição de tensores. Com recurso a um caso de estudo ilustrativo, no qual analisamos uma rede de amizade que evolui ao longo do tempo, verificou-se que a abordagem proposta apresenta várias características desejáveis, sendo concomitantemente simples, informativa, e compacta, permitindo assim a inspeção visual e a compreensão das mudanças estruturais ocorridas em redes dinâmicas, a diferentes níveis de análise.

Também contribuímos para a perspetiva prescritiva da análise de negócios através da realização de investigação aplicada em análise de decisão multicritério, no departamento de pintura de uma fábrica de montagem de automóveis. Mais especificamente, nós aplicamos a abordagem híbrida AHP-PROMETHEE, suportada pela técnica de GAIA, ao novo problema de avaliação dos méritos relativos de um conjunto de planos de pintura, com o objetivo de ajudar o decisor da fábrica de montagem a melhorar o planeamento e a otimizar o processo de pintura. Com base nos resultados do nosso caso de estudo, concluímos que a abordagem híbrida AHP-PROMETHEE é operacionalmente útil, devido à sua simplicidade e facilidade de aplicação, e ligeiramente mais fiável e robusta do que as respetivas versões autónomas (*i.e.*, o AHP e o PROMETHEE).

Palavras-Chave: Análise de negócios, análise de redes sociais, análise da evolução temporal de comunidades, modelos de decomposição de tensores, modelo Tucker3, trajetórias espácio-temporais, análise de decisão multicritério, AHP-PROMETHEE, GAIA

Abstract

Managers often have to analyse and look for relevant pieces of information in vast amounts of data in order to inform and guide their decisions. However, finding the right information in a timely manner becomes more difficult as the size of data grows, hampering the ability of managers to act and decide quickly. Business analytics is a promising approach to address the information overload faced by managers in the current Information Age. Business analytics transforms data into value by uncovering non-trivial patterns and relationships in large amounts of data using quantitative and computational techniques. The goal is to provide managers with valuable business insights that help them make effective data-driven decisions.

Business analytics encapsulates three major perspectives: the descriptive, the predictive, and the prescriptive. This thesis makes contributions on both descriptive and prescriptive analytics, which have been largely unexplored in recent literature compared to predictive analytics. Our research goal is twofold: (i) to develop new data-driven methodologies for advanced descriptive analytics that distill the value contained in historical time-stamped relational data collected by organisations and (ii) to investigate the potential of using hybrid multicriteria decision analysis approaches for tackling complex industrial problems in the context of prescriptive analytics.

We contribute to descriptive analytics by developing two complementary methodologies that extract patterns from temporal relational data, namely, (i) a methodology for monitoring dynamic communities in evolving networks at different levels of temporal resolution and (ii) a methodology for visualising evolving networks at the node-level or the community-level. The first methodology explores social network analysis measures, community detection methods, and an event-based framework for community evolution analysis, to discover evolutionary customer profiles in dynamic customer networks and detect changes in their evolution (*e.g.*, splits and merges of customer communities). We show the usefulness of the proposed methodology by applying it to an evolving customer network derived from the customer base data of a large Portuguese company. The second methodology is a new approach for visualising evolving networks that relies on displaying spatio-temporal trajectories of nodes, or communities, in a meaningful and interpretable 2D space. To generate the visualisations we explore three-order tensors and multiway analysis methods. An illustrative case study on an evolving friendship network revealed that the proposed approach has

several desirable features, being concomitantly simple, informative, and compact, thus allowing the visual inspection and understanding of structural changes in the evolution of entities in dynamic networks.

We contribute to prescriptive analytics by conducting applied research in the paint shop of an automobile assembly plant using multicriteria decision analysis. We tackled the novel problem of evaluating the relative merits of a set of vehicle painting plans using the AHP-PROMETHEE hybrid approach, aided by GAIA, with the goal of helping the decision maker of the automobile assembly plant enhance the paint shop planning and optimise the painting process. Based on the results of our case study we conclude that the AHP-PROMETHEE approach is operationally useful, due to its simplicity and ease of application, and slightly more reliable and robust than the corresponding stand-alone versions.

Keywords: Business analytics, social network analysis, community evolution analysis, multi-way analysis, Tucker3 model, spatio-temporal trajectories, multicriteria decision analysis, AHP-PROMETHEE, GAIA

Contents

List of Figures

List of Tables

List of Acronyms

AHP Analytic Hierarchy Process.
ALS Alternating Least Squares.
ANOVA ANalysis Of VAriance.
ANP Analytic Network Process.
ARD Automatic Relevance Determination.

B2B Business-to-Business.
BA Business Analytics.

CI Consistency Index.
CP CANDECOMP/PARAFAC.
CPM Clique Percolation Method.
CR Consistency Ratio.
CRM Customer Relationship Management.
CV Coefficient of Variation.

DEDICOM DEcomposition into DIrectional COMponents.
DIFFIT DIFFerence in FIT.
DM Decision Maker.
DNA Dynamic Network Analysis.

EC Energy Consumption.
ELECTRE ELimination and (Et) Choice Translating REality.

GAIA Geometrical Analysis for Interactive Aid.
GDM Group of Decision Makers.
GED Group Evolution Discovery.

HITS Hyperlink-Induced Topic Search.
HOSVD High-Order Singular Value Decomposition.

LP Label Propagation.

MADM Multi-Attribute Decision Making.
MAUT Multiple Attribute Utility Theory.
MAVT Multiple Attribute Value Theory.
MCDA Multicriteria Decision Analysis.
MCDM Multicriteria Decision Making.
MDS MultiDimensional Scaling.
MEC Monitor of the Evolution of Clusters.
MECnet Monitor of the Evolution of Communities in Networks.
MODM Multi-Objective Decision Making.

NMI Normalised Mutual Information.
NPV Number of Painted Vehicles.

OSN Online Social Networks.

PC Paint Consumption.
PCA Principal Component Analysis.
PROMETHEE Preference Ranking Organization METHod for Enrichment Evaluations.

QI Quality Index.

RCI Random Consistency Index.

SMART Simple Multi-Attribute Rating Technique.

SNA Social Network Analysis.

STATIS Structuration des Tableaux A Trois Indices de la Statistique.

SVD Singular Value Decomposition.

TOPSIS Technique for Order of Preference by Similarity to Ideal Solution.

WWW World Wide Web.

List of Symbols

$A_{n\times n}$ an adjacency matrix or sociomatrix.
$A_{n\times m}$ an incidence matrix.
$L_{n\times n}$ a Laplacian matrix.
Sk_2 population Pearson's second skewness coefficient.

b_i betweenness centrality of node i.
b_e betweenness centrality of edge e.

Cl_i closeness centrality of node i.
C global clustering coefficient of a graph.
c_i local clustering coefficient of node i.
ξ_i community structure of a network at time point t_i.
$Com_{t_i}^{m}$ community m found at time point t_i.
k_{t_i} number of communities found at time point t_i.
CR consistency ratio of the AHP method.
$\rho_{X,Y}$ population Pearson's correlation coefficient between variables X and Y.
CV coefficient of variation.

$\overline{k}$ average degree centrality of a graph.
p_k degree distribution.
k_i degree centrality of node i.
k_i^+ in-degree centrality of node i.
k_i^- out-degree centrality of node i.
k_i^w strength of node i.
ρ graph density.
D graph diameter.

ε_i eccentricity of node i.
x_i eigenvector centrality of node i.

l average geodesic distance in a graph.
l^{-1} harmonic average geodesic distance for graphs with more than one connected component.
G graph.

$d(i,j)$ geodesic distance, *i.e.* length of the shortest path between nodes i and j given by the number of edges linking both nodes.

μ population mean.
λ split threshold of the MECnet framework.
τ matching threshold of the MECnet framework.
$\widetilde{x}$ population median.
Q modularity function.

N_i neighbourhood of node i.
C_y network centralisation index.
$I(A,B)$ normalised mutual information between data partitions A and B.
n graph order or number of vertices in a graph (Chapters 2 and 4)/number of criteria (Chapter 5).
m graph size or number of edges in a graph (Chapter 2)/ number of options (Chapter 5).

s Gaussian threshold of the preference function (PROMETHEE).

q indifference threshold of the preference function (PROMETHEE).

$\phi^{-}(\mathbf{a})$ negative (or entering) outranking flow of alternative a (PROMETHEE).

$\phi(\mathbf{a})$ net outranking flow of alternative a (PROMETHEE).

$\phi^{+}(\mathbf{a})$ positive (or leaving) outranking flow of alternative a (PROMETHEE).

p preference threshold of the preference function (PROMETHEE).

R graph radius.

$A(G)$ set of arcs of a directed graph G.

$E(G)$ set of edges of graph G.

$V(G)$ set of vertices of graph G.

σ population standard deviation.

P number of components, or levels, of component matrix **A**.

Q number of components, or levels, of component matrix **B**.

R number of components, or levels, of component matrix **C**.

g_{pqr} entry of a core tensor.

$\mathscr{G}$ core tensor.

I number of entities of mode A of a tensor.

J number of entities of mode B of a tensor.

K number of entities of mode C of a tensor.

A component matrix associated with mode A of a tensor.

B component matrix associated with mode B of a tensor.

C component matrix associated with mode C of a tensor.

A row-entities mode of a tensor.

B column-entities mode of a tensor.

C fiber-entities mode of a tensor.

N order, or way, of a tensor.

x_{ijk} entry of a three-order tensor.

χ high-order tensor.

σ^2 population variance.

w length of a sliding window, as given by the number of time points.

W_k timestep of a window, *i.e.* a time interval starting at t_i and ending at t_{i+w}.

Part I

Introduction

Chapter 1

Introduction

An investment in knowledge always pays the best interest.

Benjamin Franklin

1.1 Context

Decision making is a ubiquitous human activity and a key facet of business (Allison et al., 1992; Holsapple et al., 2014). Much of the work within organisations relates to acts of problem solving and decision making and these are two key functions of managers and business executives (Simon et al., 1987). Everyday managers have to decide and these decisions address everything from day-to-day operational issues to long-range strategic planning. The ability of managers to solve problems and make decisions shape the organisations they run and there is a tight link between the quality of managers' decisions and the performance of the organisations (Blenko et al., 2010). Ultimately, an organisation's value is given by the sum of the decisions it makes and executes, being its fate largely dependent on the managerial choices and actions. Besides the pressure to make the right decisions, managers also feel the pressure to make quick decisions in order to seize windows of opportunities and stay ahead of competition. This becomes even more critical in the fast-paced complex environments in which they operate. The pressure to act quickly is one of the reasons why managers often decide and act on rules of thumb, intuition, or business experience; in contrast structured methods are still seldom applied in the decision-making process of organisations. The generalised use of non-analytical procedures in the decision-making process is highlighted by Bana e Costa et al. (1999, p. 333), who state that "people are accustomed to take hard decisions without the support of a formal methodology".

The need for making decisions arises when people face *decision problems*. A decision problem can be defined as a choice to be made between two or more courses of action in order to reach a

set of desired goals in a situation that may have uncertainty and/or imperfect information (Hastie and Dawes, 2001). The available options are evaluated based on their ability to achieve a set of objectives and according to the decision maker's system of values, preferences, and priorities (Keeney, 1982). Addressing a decision problem can be translated into (i) choosing the best option among the available ones, (ii) delimiting the subset of reasonable options, (iii) sorting the options in descending order of overall preference, or (iv) characterising the options in terms of their consequences in order to facilitate the evaluation of their relative merits and disadvantages. To make effective decisions, one must consider the broad range of options that may accomplish the desired goals, as well as define what is really relevant for analysing them (Drucker et al., 1967).

As mentioned before, business decisions can be informed by (i) intuition, hunches or gut instinct, known as the *intuitive decision making*, or by (ii) the empirical analysis of data and facts, known as the *data-driven* or *fact-based decision making*. The intuitive decision making, which constitutes the basis of most decisions in organisations, especially the small ones, relies on intuitive judgements and unconscious expertise that do not involve explicit deduction or reasoning. Its main benefits are being fast and effortless, due to its ability to recognise patterns in a fraction of a second (Kahneman, 2011). Although these quick judgements may sometimes be accurate, especially when derived from reliable and relevant information and prior experience (Burke and Miller, 1999; Thagard, 2001), and useful in novel or unfamiliar settings, where no previous prior experience or facts are available, it has been recognised that human judgements can be limited, distorted, and prone to bias (Tversky and Kahneman, 1974). Besides, there is a risk of these intuitions being based on inaccurate or irrelevant information or experience. Another drawback is that the use of intuition is problematic in group decision making, where decisions need to be made collectively and there are conflicting points of views concerning the right course of action. Decisions based on intuition can also prove to be insufficient and unsatisfactory in situations where they need to be justified to regulatory authorities, shareholders, bosses, or the general public.

Such settings require a more deliberate and analytical approach for guiding the decision-making process. The data-driven decision making comes out as an appealing alternative for these situations, because it provides a more rational and rigorous framework, supported by structured thinking and systematic procedures. In this type of decision making, the objectives and alternative courses of action are made explicit, and decisions are backed up with objective and verifiable data and facts, which increases the transparency of the decision process. The adoption of this approach acts as a means to support, justify, and increase the confidence on managers' decisions, since its outcomes are consistent over time and over people. In addition, the process underlying this type of decision making promotes learning and allows to quantify the incremental knowledge. Despite its benefits, the data-driven decision making approach is not immune to weaknesses. Besides being slower than intuitive decision making, because it involves higher cognitive effort and the collection and analysis of data, the success of the decisions is strongly reliant upon the quality of data and the

effectiveness of its analysis and interpretation. Given that both types of decision making are fallible, successful decision making should rely on a balance between fact-based and intuitive thinking, by integrating the insights provided by the analysis of data and facts with the intuition, domain knowledge, expertise, and business understanding of managers. This view is shared by several researchers, such as Simon (1987), Thagard (2001), Patton (2003), Miller and Ireland (2005), and Kahneman (2011), to name but a few, who emphasise the complementary nature of analysis and intuition in decision making, especially when important decisions are at stake.

The digital age we are living in creates a fertile ground to implement this view. With the increasing storage and computational power, the corresponding declining costs, and the widespread presence of advanced database technologies in contemporary organisations, vast amounts of detailed data about customers, suppliers, processes, operations, and the business as a whole are being produced, collected, and stored at every moment. For instance, retail companies such as Walmart handle more than one million customer transactions per hour, which translates into approximately 2.5 petabytes of data in a single day. Facebook handles more than 300 petabytes of user-generated content every day and Amazon processed an average of 306 transactions per second at their peak in 2012, which equates to roughly 26.5 million transactions per day. The wealth of data available today is truly astonishing, which poses the problem of data deluge (also referred to as data flood, data overload or data explosion). Such information overload can prove detrimental to decision makers since it becomes more difficult to find interesting or valuable pieces of information that can support their decisions. However, there is a bright side. Data have been recently deemed as the "new oil" (Acito and Khatri, 2014), being a valuable intangible corporate asset that can be translated into meaningful insights that drive the organisations' performance. The accumulation of data in companies enables the creation of a tangible "organisational memory" (Acito and Khatri, 2014), allowing to know each element of the business environment at unprecedented levels of detail. This opens exciting opportunities, as these data potentially contain hidden knowledge and relevant insights on market, competitors, customers, processes, human resources and so on, that can support key decisions and reveal non-trivial, previously unknown, and useful patterns.

In this context, several questions arise: How can decision makers cope with the data deluge problem, tame the information overload and leverage data as means to acquire competitive advantage? How can they tap the potential of these data? How can they sift through these large amounts of data and extract valuable information that can be used to inform decisions? Business Analytics (BA) appears as a feasible solution to the information overload problem faced by managers, due to its focus on the extraction of meaningful patterns from large amounts of data using quantitative and computational techniques. Business Analytics lies at the intersection of three key disciplines: business intelligence and information systems, applied statistics, and operations research and management science, having a strong quantitative nature. The Institute for Operations Research and the Management Sciences defines Business Analytics as the "scientific process of transforming

data into insight for making better decisions" (INFORMS, 2015). Although this definition is rather vague, it is regarded as the most consensual one (Holsapple et al., 2014), as it captures the *ethos* of BA. An interrelated but more complete definition is the one proposed by Evans and Lindner (2012, p.5), who view Business Analytics as the "use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions". These two definitions encapsulate three key ingredients of BA: data, analytics, and decision/action. The first ingredient is associated with information technologies, such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) and Database Management Systems (DBMS), which enable the collection, storage, and management of data from business activities. The second ingredient (analytics) refers to the portfolio of models, methods, and techniques to derive value from these data and produce visualisations that help communicate the patterns to decision makers. Analytics performed over the collected data potentially derives knowledge that guides and supports data-driven decision making and problem solving in companies, thus yielding the last ingredient: decisions or actions. Ultimately, this process turns data growth into business growth.

BA has gained momentum in the past few years. A search on Google Scholar suggests that 4170 articles have been published in Business Analytics in a time span of one year ($[2014, 2015]$), which equates roughly to one article every two hours. This is a stark contrast to what was observed ten years ago ($[2004, 2005]$), for which the same search retrieved 313 articles. This marked upswing in the number of articles reflects the recent academic popularity of BA, which extends to practice as well. According to a study by Gartner, analytics is the top technology priority of Chief Financial Officers (CFOs) (Elliot, 2012). The prominence of BA is closely related to its perceived importance and advantages, allowing organisations to (i) derive knowledge by uncovering meaningful patterns and relationships in data, (ii) provide valuable business insights to managers and executives, (iii) translate data into competitive advantage (Holsapple et al., 2014). In fact, at least two studies have found a significant association between companies' commitment to BA and data-driven decision making, and their overall performance in terms of revenue, profit, quality of the processes, and shareholder return (Brynjolfsson et al., 2011; LaValle et al., 2013). Therefore, the adoption of BA in organisations can provide a competitive edge, and is likely to contribute to smarter and more informed decisions and better business outcomes. Note however that this success is dependent on several factors, such as the quality of the data, which must be accurate, timely, and relevant; the availability of infrastructures and systems able to support and process large amounts of data; an organisational culture that fosters and facilitates data-driven decision making; and the production of human understandable visualisations that clearly and effectively convey the insights extracted from data (Philpott, 2010).

Business Analytics is commonly viewed from three major perspectives: the descriptive, the predictive, and the prescriptive (Kiron et al., 2011; Evans and Lindner, 2012).

Descriptive analytics is the best known and most used analytics in organisations (Evans and Lindner, 2012). Descriptive analytics tries to provide a comprehensive view and context for what has happened, by analysing past and current events using historical data. The main goal is then to extract meaning and understanding from data, and the analysis does not necessarily translate into explicit decisions. Descriptive analytics rather focuses on characterising, categorising, and summarising data in order to detect hidden patterns and uncover relationships among variables that better help understand the business and the reasons behind past failure and success. It also makes use of techniques for visually representing the discovered patterns or relationships, so as to foster insight and promote understanding. Examples of questions that can be answered through descriptive analytics are: Who are the most profitable customers? How did the company's sales evolved in each region? How many customers' complaints did the company received and successfully resolved? How can we segment customers to support the development of tailored marketing campaigns? To provide answers to such questions, descriptive analytics makes use of tools such as business intelligence, statistical analysis, and exploratory data mining.

Predictive analytics is currently the most popular and most frequently discussed of the three perspectives (Holsapple et al., 2014). The question that drives predictive analytics is "What could happen?" due to its focus on predicting the future based on the extrapolation of patterns or relationships found in historical data (Evans and Lindner, 2012; Praseeda and Shivakumar, 2014). This is achieved by developing predictive models using techniques such as statistical models, machine learning, predictive data mining, and time series forecasting. These models help managers determine the most probable outcome of an event, anticipate likely scenarios, and predict future events. Examples include predicting next season's demand of a specific product, predicting the chances that a loan application will default, predicting customer churn, or predicting the next product a customer will buy given its past purchasing behaviour. However, predictive analytics does not recommend actions. This task falls within the domain of prescriptive analytics.

Prescriptive analytics is still in its infancy but has been recognised by Gartner (2015) as an "innovation trigger" for organisations and one of the most promising perspectives of advanced BA. This perspective uses data to answer "What should we do?", thus implying a call for action that usually translates into business decisions. Prescriptive analytics aims to analyse how different courses of action will affect the business performance in order to point managers toward the best or most satisfactory option. It makes use of optimisation, Multicriteria Decision Analysis (MCDA) and simulation to identify, examine, and recommend a set of potential actions that best achieve the business goals and solve a given problem. Additional tools that can be harnessed for prescriptive analytics are decision and process modelling. Prescriptive analytics is usually based on the analysis of structured data complemented with semi-structured or unstructured data, such as expert judgments or decision makers' preferences. The integrated analysis of these data allows to understand the implications of the decision options and then make suggestions based on this knowledge. Examples

of questions that can be answered through prescriptive analytics are: What is the best pricing and advertising strategy to maximize revenue? Which stock options should integrate my portfolio in order to maximize the revenue and minimize the risk? What is the optimal scheduling? Which information system should the company buy?

In this thesis, we make contributions on both descriptive and prescriptive analytics, which have been largely unexplored in recent literature compared to predictive analytics.

1.2 Goals and Research Questions

In the previous section, we addressed the importance of organisational decision making and problem solving, the complementary nature of intuitive and data-driven decision making, the challenges posed by the data deluge problem, and how different perspectives of Business Analytics can be used to tackle this problem. We also emphasised the ability of BA solutions to provide a solid data-based foundation that can both support and raise managers' confidence in their decisions and actions. In this context, the main purpose of this thesis is to:

Develop novel data-driven methodologies for advanced descriptive analytics and investigate multicriteria decision analysis methods for prescriptive analytics that are able to distill the value in historical data and support timely decision making, through the detection of meaningful patterns and delivering of business insights.

The proposed methodologies aim at improving the quality and timeliness of the input to the decision-making process, by deriving value from historical data. These methodologies should, as far as possible, yield meaningful and interpretable outputs that are easy to grasp and understand, in order to enable managers to quickly absorb the relevant information and take action, without having to deal directly with complex mathematical models.

Fact- or data-driven decisions increase managers' confidence and can potentially alleviate the stress resulting from the pressure to decide quickly and accurately in face of vast amounts of complex information. However, albeit data-driven, these methodologies rely on a certain degree of human intervention and involvement. The central element is still the human, who has the ability to feel and to decide what is the most relevant pattern or the best outcome according to his/her business understanding and valuable domain knowledge.

We decompose the research goal into three research questions:

Descriptive Analytics:

RQ1: How can we model and detect changes in historical data by harnessing both the relational and the temporal nature of data?

RQ2: How can we visualise and understand the relational meaning of the detected temporal changes?

Prescriptive Analytics:

RQ3: What is the potential of using hybrid approaches in multicriteria decision analysis to aid decision makers address real-life complex decision problems?

To answer the first research question we devised a methodology to discover and monitor customer profiles in evolving customer networks. The methodology models the similarity of customers' purchasing behaviour as a network, analyses their temporal development using Social Network Analysis (SNA) techniques, uncovers customer profiles using community detection methods, and detects different types of changes, or events, in the evolution of these profiles using event-based frameworks for community evolution analysis. **(Chapter 3)**

The second research question was addressed by developing a trajectory-based visualisation methodology that focuses on capturing and semantically understanding the changes occurring in the structure of a dynamic network at multiple levels of analysis. The proposed visualisation summarises the historical network data (or relational data) using tensor decomposition models to produce a merged view of the entire historical evolution of the data, in an informative static bidimensional space. **(Chapter 4)**

To answer the third research question we applied a hybrid approach that combines the strengths of two well known multicriteria decision analysis methods, namely, the Analytic Hierarchy Process (AHP) and the Preference Ranking Organization METHod for Enrichment Evaluations (PROMETHEE), to the problem of assessing vehicle painting plans in an automobile paint shop. We also compared the results of the hybrid approach with the ones obtained using the stand-alone versions of AHP and PROMETHEE. **(Chapter 5)**

In what regards our philosophical stance, we adopted a positivist perspective for conducting the research work presented in this thesis. In terms of epistemology, we followed the objectivist approach as we believe that reality is objective, having an independent existence, detached from any consciousness, and whose meaning can be unveiled through research. Furthermore, we view the object of study as independent of our senses, existing outside the mind of the researcher, thus

adopting a realist position in terms of ontological assumptions. This philosophical position is reflected in our statements throughout the thesis, which are based on data and facts.

1.3 Thesis Contributions

Business Analytics is the unifying theme that links the two research lines addressed in this dissertation: dynamic network analysis and multicriteria decision analysis, being both of utmost importance for the field of management science. This thesis contributes to the Business Analytics field by providing advanced methodologies for descriptive analytics and by investigating a hybrid approach in the context of prescriptive analytics. In addition, we make specific contributions to the fields of social network analysis, dynamic network analysis, information visualisation, and multicriteria decision analysis.

The contributions of this thesis are briefly summarised as follows.

- A new methodology for the identification and monitoring of evolutionary customer profiles in dynamic customer networks;
- A new event-based framework for community evolution analysis, termed Monitor of the Evolution of Communities in Networks (MECnet);
- Comparison of the type of information delivered by two different time window models, the landmark window and the overlapping sliding window, in the analysis of the evolution of customer profiles;
- Application of the developed methodology in a real-world customer network;
- Modelling of evolving network data using three-order tensors embedded with structural information derived from the network;
- A new visualisation methodology for analysing the evolution of a network, at the node-level or the community-level, based on the projection of spatio-temporal trajectories in a meaningful and understandable 2D space;
- Introduction of a novel multicriteria paint shop problem, namely, the evaluation of the relative merits of the most frequently used daily painting plans in automobile paint shops;
- Application of a hybrid AHP-PROMETHEE approach to the new paint shop problem using as a case study one of Toyota's assembly plants.

1.4 Thesis Format and Bibliographical Note

For the development of the research work presented in this dissertation, we have opted for a structure that resembles the Scandinavian Ph.D. model. This is reflected in the thesis format, which is organised as a compilation of four independent scientific articles that were already published in peer-reviewed international scientific journals. These articles were further extended for the purpose of this thesis, and represent significantly improved versions of the originally published articles. To provide a context to the independent research conducted in each one of them, this introductory chapter ties these articles together by placing them in the overarching field of Business Analytics.

The main chapters that make up the body of this dissertation correspond to improved and extended versions of the following published articles:

- **Chapter 2:** Oliveira, M. and Gama, J. (2012). An overview of Social Network Analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2):99-115. [ISI, Scopus, DBLP]
- **Chapter 3:** Oliveira, M., Guerreiro, A., and Gama, J. (2014). Dynamic communities in evolving customer networks: An analysis using landmark and sliding windows. *Social Network Analysis and Mining*, 4(1):1-19. [Scopus, DBLP]
- **Chapter 4:** Oliveira, M. and Gama, J. (2013). Visualization of evolving social networks using actor-level and community-level trajectories. *Expert Systems*, 30(4):306-319. [ISI, Scopus, EBSCO, Emerald, DBLP]
- **Chapter 5:** Oliveira, M., Fontes, D.B.M.M., and Pereira, T. (2015). Evaluating vehicle painting plans in an automobile assembly plant using an integrated AHP-PROMETHEE approach. *International Transactions in Operational Research*. In Press. [ISI, Scopus, EBSCO]

Since some of these articles benefited from the collaboration with other researchers, besides my supervisors, I would like to acknowledge their contribution. The research published in the second article (Chapter 3) was jointly conducted with Américo Guerreiro. The main tasks performed by Américo Guerreiro were: (i) to collect and preprocess the customer data, (ii) to generate the dynamic customer networks in the Gephi software, (iii) to extract the SNA measures for each window step and for each window model, (iv) to act as a bridge between me and the company, (v) to help in the interpretation of both the SNA measures and the qualitative results from a business perspective. The research published in the fourth article benefited from the collaboration and experience of Teresa Pereira, who played an important bridging role between us and TOYOTA Caetano de Portugal. Teresa Pereira convened the meetings with the decision maker and supported

us during the application of multicriteria decision analysis to the paint shop problem.

During the course of my Ph.D. studies, I published additional articles in related topics and embraced opportunities to collaborate with other researchers. These originated the following (selected) publications:

- Correa, F., Oliveira, M., Alves, L. R., Gama, J. and Corrêa, P. L. (2011). Data mining applied on grain data marts. In *Proceedings of the 8th European Federation for Information Technology in Agriculture, Food and the Environment/World Congress on Computers in Agriculture (EFITA/WCCA'11)*, pages 518-527. Czech Centre for Science and Society.

- Correa, F., Oliveira, M., Alves, L. R., Gama, J. and Corrêa, P. L. (2011). Extraction of events in agribusiness data marts of Brazil using Tucker decomposition. In *Actas da XVIII Jornadas em Classificação e Análise de Dados (JOCLAD 2011)*, pages 185-188. CLAD.

- Oliveira, M. and Gama, J. (2011). Visualizing the evolution of social networks. In Antunes, L. and Sofia Pinto, H., editors, *Progress in Artificial Intelligence*, pages 476-490. Springer.

- Oliveira, M. and Gama, J. (2012). A framework to monitor clusters evolution applied to economy and finance problems. *Intelligent Data Analysis*, 16(1):93-111.

- Tork, H., Oliveira, M., Gama, J., Malinowski, S. and Morla, R. (2012). Event and anomaly detection using Tucker3 decomposition. In *Proceedings of the 2012 Ubiquitous Data Mining Workshop (UDM 2012)*, pages 8-12.

- Siddiqui, Z., Oliveira, M., Gama, J. and Spiliopoulou, M. (2012). Where are we going? Predicting the evolution of individuals. In Hollmén, J., Klawonn, F. and Tucker, A., editors, *Advances in Intelligent Data Analysis XI*, pages 357-368. Springer Berlin Heidelberg.

- Oliveira, M., Fontes, D.B.M.M., and Pereira, T. (2013). Multicriteria decision making: A case study in the automobile industry. In *Proceedings of the 2013 International Symposium on Operational Research and Applications (ISORAP 2013)*, pages 1-8. IFORS.

- Oliveira, M., Fontes, D.B.M.M., and Pereira, T. (2014). Multicriteria decision making: A case study in the automobile industry. *Annals of Management Science (Special Issue on Operations Research and Applications)*, 3(1):109-128.

- Cerqueira, V., Oliveira, M. and Gama, J. (2015). A framework for analysing dynamic communities in large-scale social networks. In *Proceedings of the 17th International Conference on Enterprise Information Systems (ICEIS 2015) - Vol. 1*, pages 235-342. INFORMS.

- Sarmento, R., Oliveira, M., Cordeiro, M., Tabassum, S. and Gama, J. (2015). Social network analysis in streaming call graphs. In *Big Data Analysis: New Algorithms for a New Society*. In Press.

Besides these scientific collaborations, which helped me widen my knowledge on social network analysis, network science, dynamic community analysis, and tensor decomposition models, I also collaborated in the organisation of scientific conferences in the fields of machine learning and data mining, both as publicity chair and social media chair. In addition, I acted as referee in several conferences and journals in artificial intelligence, machine learning, data mining, and statistics. I also co-supervised Master students in the topics of dynamic network analysis and multicriteria decision analysis.

1.5 Thesis Outline

The present thesis is the result of four journal articles and is organised in six chapters. This introductory chapter (Chapter 1) provides the general context, motivation and purpose of our research, specifies the main research questions, summarises the original contributions of this thesis and gives an overview of the thesis organisation. A bibliographical note detailing the publications that emerged from this doctoral project and related research work is also presented.

The four chapters that constitute the main body of this dissertation are divided in two parts: Part II and Part III. Both parts fall within the broader scope of Business Analytics but each embraces a different BA perspective and pursues a distinct research line, characterised by different objectives and methodologies.

Part II encompasses three chapters (Chapters 2, 3 and 4) and its foundation is laid on network science and social network analysis. Chapter 2 provides an overview of the fundamental terms, concepts, definitions, measures, methods, and findings of network research in order to familiarise the reader with the topics addressed in the subsequent chapters. Two novel dynamic network analysis methodologies for advanced descriptive analytics are introduced in Chapters 3 and 4: one for monitoring the evolution of groups of customers exhibiting similar purchasing behaviour, and another for visualising and understanding temporal changes in the evolution of dynamic networks.

Part III comprises Chapter 5 and explores multicriteria decision analysis in the context of prescriptive analytics. Here, we introduce a novel multicriteria decision problem, identified in a real-world industrial context, and investigate a hybrid multicriteria decision analysis approach for guiding the corresponding decision-making process.

Note that these four chapters are structured as scientific articles and, thus, the specific motivation, goals, literature review, methodology, data, limitations, conclusions and future work are separately provided for each one of them. Although independent, some of these chapters are related to each other, namely, Chapters 2, 3, and 4 and, thus, there is some overlap in content. This is a consequence of the self-containedness of each chapter. Moreover, since these chapters represent improved and extended versions of the original journal articles, for the sake of clarity, we highlight the connection between them by explicitly referring to the preceding or subsequent chapter(s).

The final chapter (Chapter 6) concludes the thesis by answering the proposed research questions, reviewing the most relevant research outcomes and outlining possible future developments.

Part II

Dynamic Network Analysis for Descriptive Analytics

Chapter 2

An Overview of Network Research

The analysis of networks has attracted a burst of attention over the last decade. Two relevant reasons are: (i) the unprecedented availability of large volumes of network data, boosted by the proliferation of social media websites and the increased computing power, and (ii) the intuition that an entity's connections can yield richer information than its isolate attributes. The analysis of a wide diversity of real-world problems through the network perspective enabled important discoveries about the nature of networks arising in contexts as diverse as the social, biological and technological. Key findings of network research include the strength of the weak ties, the small-world effect, the power-law degree distributions, the community structure, the transitivity, and the homophily in networks. These characteristics of networks helped to better understand how the underlying complex systems function and behave, thus allowing the simulation and prediction of processes taking place inside these systems, such as spread of diseases, information diffusion and the emergence of power. This chapter attempts to provide a general and succinct overview of the field of social network analysis and network science in general, by focusing on its most important concepts, definitions, tasks and methods.

2.1 Introduction

The last few years have been marked by a renewed interest in the study of networks, which emerged as an invaluable abstraction for describing and quantifying complex systems in many branches of science (*e.g.*, sociology, anthropology, physics, biology, information technology, economics, management) giving rise to an extensive and burgeoning body of research. This popularity surge is a reflection of a general shift, beginning in the second half of the 20th century, away from the individualist, essentialist, and atomistic explanations toward more relational, contextual, and systemic understandings of phenomena (Borgatti and Foster, 2003). This paradigm shift is anchored in a network perspective, which views complex systems as sets of interrelated entities. According

to this perspective, social actors are embedded within networks of interconnected relationships that provide opportunities and constraints on their behaviour. In fact, the network perspective puts more emphasis on the structure of the underlying network than in the entities (*e.g.*, individuals, groups, organisations, countries, customers, web pages, proteins) themselves, assuming that the intricate web of relationships among entities determines their decisions and outcomes, as well as the behaviour and function of the overall system.

Historically, networks were first studied within the scope of graph theory, a branch of discrete mathematics whose foundations were laid out in 1736, when the Swiss mathematician Leonhard Euler published a paper with the solution to the Königsberg bridge problem. This problem consisted in finding a single trip that crossed every one of the seven bridges over the river Preger in the city of Königsberg exactly once, with the additional requirement that the trip had to end in the same place where it started. Euler discovered that such path did not existed resorting to the concept of *graph*. Since then, and especially in the 20th century, graph theory has witnessed many developments and *graphs* have been widely used as a model to represent networks of all kinds. Social scientists are believed to have been the first to explicitly incorporate network thinking as an approach to study social structures, and to adopt graph theory as a formal language to study social phenomena. The prolific investigation of networks in the social sciences, supported by the graph theory formalism, led to the development of an important research field today known as Social Network Analysis (SNA).

Social Network Analysis is a quantitative methodology mainly developed by sociologists, anthropologists, and social psychologists that has recently benefited from the collaborative efforts of academic researchers from different scientific areas (*e.g.*, physics, mathematics, computer science, biology). The origins of SNA as a basis for developing useful sociological concepts and studying social structures are difficult to discern since the field emerged as the confluence of three intellectual traditions (Berry et al., 2004): (i) the sociometric analysis developed by Jacob L. Moreno (1934) and Lewin et al. (1936) in the 1930s to study the influence of small groups dynamics on individual perceptions and behaviour, (ii) the studies on informal interpersonal relations conducted in the 1930s by the Harvard structuralists, and inspired by the work of the social anthropologist Radcliffe-Brown, and (iii) the research conducted by the Manchester anthropologists in African tribal societies, rural areas and small towns (Barnes, 1954). These traditions built on the classic works of Max Weber, who placed social action as the main pursuit of sociology, and Georg Simmel, who emphasised the formal properties of social interaction and the need to investigate the configurations that emerge from the interweaving of social relations between individuals. Although the first ideas can be traced back to the early 1930s, only in the 1960s/1970s these research traditions were consolidated by the hand of Harrison White and colleagues to give rise to the contemporary Social Network Analysis (White, 1963; Granovetter, 1973, 1974).

Since the inception of SNA, and until quite recently, the vast majority of existing network

research relied on one-shot self-reported data mostly collected using direct questionnaires, where individuals were asked to detail their interactions and relationships with others. One of the limitations of this data collection process was the production of very small datasets. As a consequence, traditional studies usually entailed some problems such as inaccuracy, subjectivity, and lack of generality due to the small sample size. Besides, traditional research was mainly concerned with how the involvement and position of a specific actor (the *ego*) in a network affected his/her outcomes and decisions (also known as the *egocentric approach*), rather than in understanding how the structure and processes taking place in the whole network explain certain outcomes (also known as the *sociocentric approach*). This trend is apparent, *e.g.*, in the studies of Granovetter (1973, 1974) on the strength of weak ties in the labour market, and in the studies of Burt (1993) on social capital and structural holes. However, the increase in computing power and storage capabilities, and the advent of disruptive technologies (*e.g.*, Web 2.0) has provided researchers with sheer amounts of highly detailed network data, as well as with the computational resources to process and analyse these data. The size of the networks available today may easily scale to millions or billions of nodes, thus enabling the investigation of the topological features and the discovery of non-trivial patterns in a wide diversity of networks. As a consequence of this evolution and of the unprecedented access to large scale networks, the research endeavour has moved from the analysis of ego-networks to whole networks. This change in research focus and the recent surge of interest in network analysis were partly triggered by the seminal papers of Watts and Strogatz (1998) on small-world networks and Barabási and Albert (1999) on scale-free networks. These two papers inspired a plethora of new research directions, which have been mainly pursued by the physicists and computer scientists, and fostered a renewed attention on networks. The recent contributions and findings from different research areas (sociology, economics, anthropology, management, political science, physics, computer science, among others) are laying the groundwork for the emergence of a unified research field commonly referred to as the *network science*. In this chapter, we offer an overview of this field by providing a short introduction to the key concepts, tasks, and methods in social network analysis and network science in general.

The remainder of this chapter is organised as follows. We begin by defining what a network is and the different types of networks we can find in the real world. Then, we explain how a network can be represented. After, we introduce the most popular measures to analyse them according to two levels of analysis: the node-level (or actor-level) and the network-level. Next, we briefly explain how link analysis can be used to identify influential and authoritative nodes. Then, we distinguish two important network models and introduce some of the most important properties of real-world networks. Later, we devote a section to the problem of finding communities in networks. This overview ends with a summary and conclusions.

Table 2.1: Network terminology for different fields of knowledge.

Mathematics	Computer Science	Social Sciences	Physics	Others
Vertex/Vertices	Node	Actor/Agent	Site	Dot
Edge	Link/Connection	Relational tie	Bond	Arc

2.2 Concept and Types of Networks

The term *social network* was first coined by the social anthropologist John Barnes in 1954 (Barnes, 1954). A social network is a set of social entities, such as people, groups or organisations, who interact and are socially related to each other. The established relationships can be of personal or professional nature and can range from casual acquaintance to close familiar bonds. Besides social relations and interactions, links in a social network can also represent financial or instrumental assistance, similarity, flow of information etc. The relationships and interactions among individuals are abstract concepts that can be inferred, for instance, from verbal or written communication (*e.g.*, phone calls, email exchanges, chatrooms), physical or virtual contiguity (*e.g.*, participation in the same events, membership of the same professional groups, visiting the same websites, belonging to the same fitness club or the same school class, living in the same neighbourhood), scientific collaboration (*e.g.*, co-authoring a publication) or explicit links in Online Social Networks (OSN) platforms (*e.g.*, Facebook, Linkedin). In a broader sense, a network is constructed from relational data, and can be defined as a collection of entities and the relationships defined on them. These networks are usually abstracted as mathematical graphs where vertices represent the entities and edges represent the relationships. The structure of these networks is determined by these relationships and constitutes the object of study of network research. Social Network Analysis offers a set of concepts, methods, techniques, and tools to empirically investigate such network structures, aiming to understand the relations among the entities and the implications of these relationships. Being an interdisciplinary endeavour, applications of SNA can be found in a myriad of branches, ranging from bibliometrics, epidemiology, biology, marketing, economics, finance, sociology to crime prevention, just to name a few. Since networks have been studied independently by distinct disciplines, each one developed its own jargon. To avoid ambiguity and clarify the adopted language, in Table 2.1 we present the network terminology used in different fields. Throughout this thesis we will use these terms interchangeably.

At the heart of the SNA approach lies the assumption that an actor's relationships carry more value than his/her individual characteristics in the process of explaining social outcomes. This idea reflects a fundamental axiom in network analysis, which is the notion that entities are not independent but rather influence each other (Borgatti and Li, 2009). The study and understanding of such mechanisms sheds light on the behaviour of the social systems that generated those networks.

The main goal of SNA is, therefore, to examine both the contents and patterns of relationships in social networks in order to understand the relations among actors and the implications of these relationships.

Common tasks of Social Network Analysis involve (i) the identification of key players in the network, *i.e.*, influential, prestigious or central actors, using specific measures; (ii) the identification of hubs and authorities, using link analysis algorithms, and (iii) the discovery of communities, using community detection techniques. These tasks are relevant not only to study research problems that help advance the scientific knowledge on a diversity of fields, but also to derive value from networks in business-driven applications. For instance, companies can use SNA to maximize positive word-of-mouth of their products by targeting the customers with the highest network value (Richardson and Domingos, 2002; Domingos, 2005; Leskovec et al., 2007a), or use information regarding a user's position in the network to devise personalised marketing campaigns. Mobile network operators can also leverage data on the communication among subscribers (better known as phone call networks) to identify customer profiles and recommend personalised mobile phone tariffs according to these profiles. Another application of network analysis in telecommunication companies is churn prediction. By detecting changes in the patterns of call activity it is possible to detect customers who may potentially switch to a competitor (Wei and Chiu, 2002; Dasgupta et al., 2008). SNA can also be applied to networks of organisational communication (*e.g.*, Enron company dataset) to uncover patterns of communication among employees and managers based on the analysis of the frequency and direction of formal/informal email communication. These patterns can help identify people engaged in fraudulent activities (Shetty and Adibi, 2004; Xu and Chen, 2005) or support the formation of project teams (Reagans et al., 2004).

Besides social networks, there are other types of structures in the real world that can be modelled as networks. According to Newman (2003b), real-world networks can be categorised into four main types: social networks, information networks (or knowledge networks), technological networks, and biological networks.

As previously mentioned, social networks are the ones that arise as a result of human socialisation and interaction. Examples include studies of friendship networks (Van De Bunt et al., 1999), informal communication networks within companies (Ritter, 1999), collaboration networks (*e.g.*, networks of co-appearance in movies by actors, in which two actors are connected if they appeared together in a movie, and networks of co-authorship among academics, in which individuals are linked if they co-authored one or more papers) (Newman, 2001), interlocking directorship networks (where nodes represent corporations and there is a corporate link if they share one or more directors) (Levine, 1972) and networks of inter-organisational partnerships (where nodes are organisations and links represent partnership ties among them) (Powell et al., 1996).

In turn, information or knowledge networks model the exchange of information among entities usually aiming to enhance knowledge diffusion, business, or social aims. Examples include peer-

to-peer networks, networks of citations between academic papers, commonly represented by an acyclic-directed graph where vertices represent papers and there is a direct edge if paper A cites paper B; and bipartite networks of users-objects (Zhou et al., 2007), which are generally modelled through bipartite graphs and may represent consumers' explicit preference for a set of commercial products or services (*e.g.*, books, travel destinations). Another example of an information network, and perhaps the most extensively studied, is the World Wide Web (WWW), where vertices represent static web pages and directed edges correspond to the hyperlinks between them (Kleinberg, 1999; Broder et al., 2000).

Technological networks are man-made networks designed for the distribution of some commodity or resource (*e.g.*, electricity, information). Examples include networks of roads and railways, power grids, networks of airline routes, and networks of physical connections between computers (Internet).

The last type of networks are the so-called biological networks (Alon, 2003) and, as the name implies, are those that arise from biological processes, such as networks of chemical reactions among metabolites, protein interaction networks, genetic regulatory networks, real neural networks, and food webs or predator-prey networks.

2.3 Representation of Networks

The graph theory framework provides an appropriate set of concepts, models, and tools for the exact mathematical treatment of networks. Therefore, networks are formally represented as graphs.

A graph is composed of two fundamental units: vertices and edges. Every edge is defined by a pair of vertices, also called its endpoints. Vertices can represent a wide variety of tangible and intangible entities (*e.g.*, people, organisations, countries, computers, publications, products, words, capital stocks, brain neurons, plants and animals) contingent on the application field. In turn, an edge is the line that connects two vertices and, likewise, it can represent numerous kinds of relationships between individual entities (*e.g.*, communication, cooperation, friendship, partnership, kinship, acquaintances, interaction, similarity, and trade). Edges may be weighted or unweighted, and directed or undirected, depending if the nature of the relation is asymmetric or symmetric.

Formally, a graph G consists of a non-empty set $V(G)$ of vertices and a set $E(G)$ of edges, being defined as $G = (V(G), E(G))$. According to Diestel (2005), the order of a graph G is given by the total number of vertices n or, mathematically, $|V(G)| = n$. Likewise, the size of a graph G is the total number of edges $|E(G)| = m$. The maximum number of edges in a graph is $m_{max}(n) = \frac{1}{2}n(n-1)$, for undirected graphs, and $m_{max}(n) = n(n-1)$, for directed ones. If self-loops (*i.e.*, edges that connect a vertex to itself) are allowed, it is necessary to add up n to these formulas, thus obtaining $m_{max}(n) = \frac{1}{2}n(n+1)$ for undirected graphs, and $m_{max}(n) = n(n+1)$ for directed graphs.

There are two conventional types of graph-theoretic data structures for graph representation:

(i) list structures and (ii) matrix structures. These data structures are appropriate for graph storage and allow their further analysis using automatic tools. List structures, such as incidence lists (give the edges incident on each node), adjacency lists (give the direct neighbours of each node), and edge lists (give the pairs of adjacent nodes), are suitable for storing sparse graphs because they only keep information about links that are present in the network and, thus, are more efficient data structures. On the other hand, matrix structures, such as incidence matrices ($A_{n\times m}$), adjacency matrices or sociomatrices ($A_{n\times n}$), Laplacian matrices ($L_{n\times n}$, with the diagonal entries indicating the degree of each node and the off-diagonal entries containing information on the adjacency of nodes), and distance matrices (entries of the matrix are the lengths of the shortest paths between pairs of vertices) are appropriate to represent full or dense matrices. In an adjacency matrix, self-loops are represented in the main diagonal.

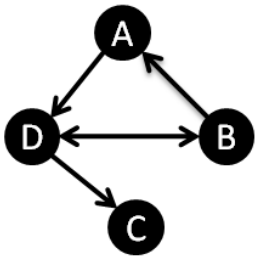
Several types of graphs can be used to model different kinds of networks. For instance, graphs can be classified according to the direction of their links. This leads us to the differentiation between undirected and directed graphs. Undirected graphs (or undirected networks) are graphs whose edges connect unordered pairs of vertices or, in other words, each edge of the graph concomitantly connects two vertices. A stricter type of graph is the so-called directed graph (or directed network). Directed graphs, or in the abbreviation form *digraphs*, can be defined as graphs whose edges have an orientation assigned to them (these edges are also termed arcs), so the order of the vertices they link matters. Formally, a directed graph is an ordered pair $(V(G),A(G))$ consisting of a non-empty set $V(G)$ of vertices and a set $A(G)$ of arcs, disjoint from $V(G)$, of arcs. If e_{12} is an arc, and v_1 and v_2 are vertices, such that $e_{12} = (v_1, v_2)$, then e_{12} is said to join v_1 to v_2, being the first vertex v_1 called the initial vertex, or tail, and the second vertex v_2 called the terminal vertex, or head. Graphically, directed edges are depicted by arrows, indicating the direction of the linkage. This type of graphs can be either cyclic, *i.e.*, graphs containing closed loops of edges or "ring" structures, or acyclic (*e.g.*, trees). A typical example of an undirected graph is Facebook because in this OSN the established friendship tie is mutual, reciprocal or two-sided (*e.g.*, if an individual accepts a friend request from a given person then it is implicitly assumed that they are friends with each other). Similarly, the micro-blogging service Twitter is an example of a directed graph because a user can be followed by others without necessarily following them. In this case, the tie connecting a pair of users is directed, with the tail being the follower and the head being the followed, meaning that a one-way relationship is established.

Regarding the values assigned to edges, we can make a distinction between unweighted and weighted graphs. Unweighted graphs are binary since edges are either present or absent. In weighted graphs each edge is associated with a weight, *i.e*, a positive real number $w \in \mathbb{R}_0^+$ that represents the strength of the relationship being modelled. According to Mark Granovetter (1973, 1974), in social networks the weight of a tie is generally a function of duration, emotional intensity, frequency of interaction, intimacy, or exchange of services. Therefore, strong ties usually represent

close friends whereas weak ties typically represent acquaintances. In other kinds of networks, the weight of an edge can represent a variety of things usually associated with similarity, distance, capacity, time, or frequency (*e.g.*, number of phone calls or messages, number of citations, number of co-authored papers, travel distance between cities, degree of similarity among entities or the number of exchanged products).

For undirected and unweighted graphs, adjacency matrices are binary (as a consequence of being unweighted) and symmetric (as a consequence of being undirected, meaning that $a_{ij} = a_{ji}$), with $a_{ij} = 1$ indicating the presence of an edge and $a_{ij} = 0$ denoting the absence of an edge between vertices i and j. For directed and weighted graphs, the entries of such matrices take values from the interval $[0, w_{max}]$ and are non-symmetric. In both cases, we deal with non-negative adjacency matrices.

Vertex	A	B	C	D
A	-	0	0	1
B	1	-	0	1
C	0	0	-	0
D	0	1	1	-



Vertex	Vertex
A	D
B	A
B	D
D	B
D	C

Figure 2.1: A directed and unweighted graph G represented by means of an adjacency matrix (left-hand side of the figure) and an edge list (right-hand side of the figure).

In Figure 2.1, we provide an example of how a network can be represented by an adjacency matrix, by a graph model, and by an edge list.

2.4 Classical Measures

In this section, we present classical centrality measures used in network analysis. These measures are useful as they provide a summary description of the network topology and enable the identification of the most central nodes.

The measures we will introduce are classified according to the level of analysis one wants to perform, which can be: (i) at the level of basic units, *i.e.*, nodes, or (ii) at the level of the whole network. The former explores general measures of centrality as a way to understand the position of a vertex within the overall structure of the graph, thus helping to identify key players in the network. The latter provides higher-level information of the network, allowing for a description of its overall structure and offering clues about properties of the underlying graph.

2.4.1 Node-Level Measures

One of the most relevant tasks in SNA is the computation and interpretation of centrality measures. Centrality, or prestige, is often used for finding important nodes by analysing their position within the overall structure of the network. The concept of *importance* is multidimensional. Each relevant dimension, such as reachability, embeddedness, influence, support, ability to span structural holes, and control the information flow, is captured by different centrality measures. Examples of classical measures include the degree centrality, the betweenness centrality, the closeness centrality, and the eigenvector centrality. The first three were proposed by Freeman (1979) and were originally devised for unweighted networks. More recently, Opsahl et al. (2010) came up with extensions to weighted networks. The fourth measure - eigenvector centrality - was later proposed by Bonacich (1987) and has its foundations on spectral graph theory. It became especially popular after being used as the basis of the well known Google PageRank algorithm, which we will introduce in the next section.

Although other node-level measures were proposed in the literature (*e.g.*, the integration and radiality measures of Valente and Foreman (1998)), in this subsection we will focus on explaining the classical ones. These measures determine the relative importance of a node within the network, showing how the relationships are concentrated in a few individuals. High values of the centrality measures are associated with powerful and influential actors in the network because their central position offers them several advantages, such as easier and quicker access to other actors in the network (useful for accessing resources such as information) and ability to exert control over the flow between the other actors (Freeman, 1979). These central actors are also called “focal points”. At the end of the section we will also introduce the concept of transitivity and explain how it can be computed using a clustering coefficient.

Note that some of these node-level measures (*e.g.*, degree centrality, betweenness centrality, and closeness centrality) may need to be normalised in order to perform comparisons of networks with different orders and sizes.

Degree Centrality

The degree centrality, or valency, of a node i, usually denoted as k_i, is a measure of the immediate adjacency and the involvement of the node in the network. It is computed as the number of edges incident on a given node or, likewise, as the number of direct neighbours or connections of node i. The neighbourhood N_i is thus defined by the set of nodes that are directly connected to i. Thus, the node degree can be computed in at least two different ways: based on the adjacency matrix or based on the neighbourhood of a node. In Equations (2.1) and (2.2) we present each one of the alternatives, for undirected networks. Despite its simplicity, degree is an effective measure to assess the local importance and direct influence of a node in a social network. Yet, it has some limitations.

The main one being that it does not take into consideration the global structure of the network and, thus, the extent of a node's influence in the whole network.

$$k_i = \sum_{j=1}^{n} a_{ij}, \qquad 0 < k_i < (n-1), \tag{2.1}$$

where a_{ij} is the entry of the ith row and jth column of the adjacency matrix $\mathbf{A}$.

$$k_i = |N_i|, \qquad 0 < k_i < (n-1), \tag{2.2}$$

where $|N_i|$ is the size of the neighbourhood of node i.

For directed networks, there are two variants of degree centrality: the in-degree centrality, denoted as k_i^+, and the out-degree centrality, denoted as k_i^-. The former is given by the number of incoming nodes (*i.e.*, number of edges targeting vertex i) and the latter by the number of outgoing nodes (*i.e.*, number of edges emanating from vertex i), as defined in Equations (2.3) and (2.4).

$$k_i^+ = \sum_{j=1}^{n} a_{ji}. \tag{2.3}$$

$$k_i^- = \sum_{j=1}^{n} a_{ij}. \tag{2.4}$$

On weighted networks, strength k_i^w is the equivalent of degree centrality, being computed as the sum of the weights of the links connected to a given node i, as expressed by Equation (2.5).

$$k_i^w = \sum_{j=1}^{n} a_{ij}^w, \tag{2.5}$$

where a_{ij}^w is the entry of the ith row and jth column of the weighted adjacency matrix $\mathbf{A^w}$.

Perhaps more importantly than analysing the degree centrality of nodes in a network, is the analysis of its degree distribution. The degree distribution p_k is a relevant network property which indicates for each possible degree value k the fraction of nodes in the network having this degree centrality value. Formally, it is a probability distribution given by $p(k) = \frac{n_k}{n}$, where n_k is the number of nodes in the network with degree k and n is the order of the network. In the past few years, a significant research endeavour has been devoted to the study of the degree distribution of different types of real-world networks, which turned it possible to classify a network based on this distribution. The scientific interest in the degree distribution was spurred by the seminal work of Barabási and Albert (1999); Barabási and Bonabeau (2003), who discovered that the distribution of the degree centrality of many real-world networks of different nature (*e.g.*, social networks, protein interaction networks, the World Wide Web network) follows a power-law distribution

(at least asymptotically), a finding that has been corroborated by a striking number of studies (Huberman and Adamic (1999); Eisenberg and Levanon (2003); Jeong et al. (2003); Palla et al. (2005); de Blasio et al. (2007); Stephen and Toubia (2009); Clauset et al. (2009); Gabaix (2009); Petersen et al. (2011), just to name a few). This means that the distribution of the node degree is very heterogeneous and highly right-skewed, with a large majority of vertices having a low degree centrality and a small number having a high degree centrality. Networks exhibiting this property are known as *scale-free networks*, an expression coined by Barabási and Albert (1999). Other common functional forms of the degree distribution are exponential (*e.g.*, railways and power grids networks) and power-law with exponential cut-offs (*e.g.*, networks of movie actors and some collaboration networks). This topic is discussed further in Section 2.6.

Betweenness Centrality

Betweenness centrality quantifies the importance of a network element (node or edge) based on the frequency of its occurrence in the shortest paths linking all possible pairs of nodes in the network. The intuitive idea behind this measure is to identify graph elements that act as structural *bridges*, *i.e.*, that connect dense regions of the network and without which the information would not pass from a region to the other. An edge with high betweenness centrality is likely to act as a bridge between such regions because it appears in many shortest paths. In the context of SNA, structural bridges are, *e.g.*, connections outside an individual's circle of close friends. These connections are of great interest for individuals seeking to access new information and resources, since they ease the diffusion of information across entire communities (Kossinets and Watts, 2006). However, situations like these are quite rare in real-world scenarios and, even if they happen, the advantages they confer are usually temporary, due to the temporal instability of such edges. A more common and realistic situation is the presence of *local bridges*, which correspond to the shortest route between different network regions. However, unlike bridges, they do not represent the only possible path by which information can flow from one region to the other, and their removal does not disconnect parts of the network. Still, the removal of local bridges increases the distance that a node from one region needs to traverse to reach another network region. Nodes with high betweenness centrality are usually located at the ends of those edges and are known as *gatekeepers*. These nodes occupy a strategic position in the network structure, which allows them to control the information transfer between different network regions, either by blocking the information between them or by accessing it before other nodes belonging to their region.

Node betweenness b_i is computed as the fraction of shortest paths that go through a given node among all shortest paths in the network (Equation (2.6)). This is a global centrality measure since it requires complete information about the network in order to compute all pairs shortest paths. Since it is based on the computation of shortest paths for the whole network, betweenness centrality is computationally demanding.

$$b_i = \sum_{j,k \in V(G) \setminus i} \frac{\sigma_{jk}(i)}{\sigma_{jk}}, \tag{2.6}$$

where σ_{jk} denotes the number of shortest paths between vertices j and k (usually $\sigma_{jk} = 1$), and $\sigma_{jk}(i)$ expresses the number of shortest paths passing through node i.

Edge betweenness centrality b_e is computed in a similar way, with the difference that now we are interested in the fraction of shortest paths that run along a given edge of the network (Equation (2.7)).

$$b_e = \sum_{i,j \in V(G)} \frac{\sigma_{ij}(e)}{\sigma_{ij}}, \tag{2.7}$$

where $\sigma_{ij}(e)$ expresses the number of shortest paths between nodes i and j that run through edge e. The sum indicates that this fraction needs to be computed for every pair of nodes i and j in the network.

Closeness Centrality

Closeness centrality Cl_i quantifies the importance of a node based on its ability to reach other nodes in a network through shortest paths. It measures how many steps on average it takes for an actor to reach everyone else in the network. The higher the closeness centrality, the less the cost for a node to reach the rest of the network. In a social network, actors with high closeness centrality can quickly spread information to, or get in touch with, other actors. Closeness centrality is computed as the inverse of the sum of the shortest path distances between a node i and the remaining $n-1$ nodes in a network of order n. This measure is only computed for nodes within the largest component of the network (or at least within connected components), using the following equation:

$$Cl_i = \frac{n-1}{\sum_{j \in V(G) \setminus i} d(i,j)}. \tag{2.8}$$

Eigenvector Centrality

Eigenvector centrality x_i assigns a relative score to each node that measures how well a given node is connected to other well connected nodes. This score is given by the first eigenvector of the adjacency matrix. The basic idea behind eigenvector centrality is that the power and status of a node is recursively defined by the power and status of its direct connections. In other words, the centrality of a given node i is proportional to the sum of the centralities of i's neighbours. As a result, nodes connected to highly central nodes have greater centrality than those connected to less central ones. This is the assumption behind the eigenvector centrality, which can be defined as the

solution for the following equation:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} a_{ij} x_j, \tag{2.9}$$

where x_i/x_j denotes the centrality of vertex i/j, a_{ij} represents an entry of the adjacency matrix $\mathbf{A}$ ($a_{ij} = 1$ if vertices i and j are connected by an edge, and $a_{ij} = 0$ otherwise) and λ denotes the largest eigenvalue of $\mathbf{A}$.

Eigenvector centrality is a more elaborated version of the degree, once it assumes that not all connections have the same importance by taking into account not only the quantity, but especially the quality of these connections.

Local Clustering Coefficient

Social networks are naturally transitive (Wasserman and Faust, 1994; Newman and Park, 2003), which means that a given actor's friends are also likely to be friends with each other. The transitivity property can be quantified by a clustering coefficient that can be global, *i.e.*, computed for the whole network, or local, *i.e.*, computed for each node. Watts and Strogatz (1998) proposed a local version of the clustering coefficient, denoted c_i. In this context, transitivity is a local property of a node's neighbourhood that indicates the level of cohesion between the neighbours of a node. This coefficient c_i measures the propensity of connection between node pairs that share a mutual neighbour, and is computed as the observed number of links between the neighbours of a given node divided by the maximum possible number of links between them. In Equation (2.10) we present the coefficient for undirected graphs and in Equation (2.11) we provide the modified coefficient for directed ones. Note that this coefficient, as defined here, can only be computed for nodes with at least one neighbour (*i.e.*, $N_i \neq \emptyset$).

$$c_i = \frac{2|\left\{e_{jk} : j,k \in N_i, e_{jk} \in E\right\}|}{k_i(k_i-1)}, \tag{2.10}$$

$$c_i = \frac{|\left\{e_{jk} : j,k \in N_i, e_{jk} \in A\right\}|}{k_i(k_i-1)}, \tag{2.11}$$

where N_i is the neighbourhood of vertex i, vertex j and vertex k are neighbours of vertex i, e_{jk} represents the edge that connects vertex j to vertex k, k_i is the degree centrality of node i, and $|\left\{e_{jk}\right\}|$ indicates the observed number of links established between the neighbours of vertex i. By definition, $0 \leq c_i \leq 1$, where 0 indicates low clustering in the neighbourhood of node i and 1 indicates a fully transitive node's neighbourhood (*e.g.*, in an acquaintance network this means that every friend of actor i knows each other).

2.4.2 Network-Level Measures

Before explaining each one of the network-level measures, there are three fundamental concepts that should be first introduced: path, geodesic distance between two vertices, and eccentricity of a vertex.

A path is a sequence of vertices in which consecutive pairs of non-repeating vertices are linked by edges; the first vertex of a path is called the *start vertex* and the last vertex of the path is called the *end vertex*. In unweighted graphs, the length of a path is given by the number of edges in that path. The concept of *paths* underpins more complex concepts, such as the *connected component*. A connected component is a connected subgraph, where for any pair of vertices there is at least one path going from one vertex to another. More specifically, a component is a connected region in the graph, such that (Easley and Kleinberg, 2010):

1. Every vertex in the subgraph has a path to every other;
2. The subgraph is a free-standing piece of the graph, not a connected part of a larger piece.

For instance, the graph of Figure 2.1 has only one connected component.

Of particular interest is also the concept of geodesic distance between nodes i and j, denoted as $d(i,j)$. The geodesic distance is the length of the shortest path, or the minimal path, between a pair of nodes i and j, as given by the number of traversed edges (in unweighted graphs).

In turn, the eccentricity ε_i is the greatest geodesic distance between a given vertex i and any other in the graph, as defined in the following equation:

$$\varepsilon_i = \max_{j \in V(G) \setminus i} d(i,j). \tag{2.12}$$

These concepts underpin some of the network-level measures that will be introduced, namely, the diameter/radius and the average geodesic distance.

Diameter and Radius

The diameter D is a measure of reachability in the network, which is defined as the maximum eccentricity in the network, *i.e.*, the longest of all the shortest paths (Equation (2.13)). Likewise, the radius R is the minimum eccentricity in the network, *i.e.*, the shortest of all the shortest paths (Equation (2.14)). Sparser networks have generally greater diameter than complete networks, due to the existence of fewer paths between pairs of nodes. The smaller the diameter is, the shorter is the upper bound for the expected path between any arbitrary pair of nodes. Leskovec et al. (2005) found evidence that, for certain types of real-world networks, the effective diameter shrinks over time. In the context of SNA, the diameter gives an idea of the proximity of pairs of actors in the network, indicating how far two individuals are, in the worst of cases.

$$D = \max\{\varepsilon_i : i \in V\}. \tag{2.13}$$

$$R = \min\{\varepsilon_i : i \in V\}. \tag{2.14}$$

Average Geodesic Distance

The average geodesic distance l, also known as the average path length, measures how far apart nodes are, on average, in the network. In the SNA context, the average geodesic distance is an useful indicator of the efficiency of information flow within the network. In undirected networks, this measure is computed according to Equation (2.15).

$$l = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} d(i,j), \tag{2.15}$$

where $d(i,j)$ is the geodesic distance between nodes i and j, and $\frac{1}{2}n(n-1)$ is the number of possible edges in an undirected network comprising n nodes.

For networks with more than one connected component, the previous formula does not hold, because the geodesic distance is conventionally defined as infinite when there is no path connecting two vertices. In such situations, it is more appropriate to use the harmonic average geodesic distance l^{-1}, defined in Equation (2.16) for undirected networks, since it turns infinite distances into zero, thus nullifying their effect on the sum.

$$l^{-1} = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i \geq j} \frac{1}{d(i,j)}. \tag{2.16}$$

Note that, for both cases (Equations (2.15) and (2.16)), if the corresponding undirected network allows self-loops, the first denominator should be replaced by $\frac{1}{2}n(n+1)$ so as to include the distance from each node to itself in the computation of the average.

Average Degree Centrality

The average degree centrality $\overline{k}$ measures the average network connectivity and indicates the average number of connections a node has in the network. It is computed as the average of the degrees of all network nodes, as given by the following equation:

$$\overline{k} = \frac{1}{n} \sum_{i=1}^{n} k_i. \tag{2.17}$$

Reciprocity

Reciprocity r is a specific quantity for directed networks that measures the tendency of pairs of nodes to form mutual connections. There are several ways to compute this measure. The most popular and intuitive way is to compute the fraction of mutual connections in the network, as shown in Equation (2.18). The value of reciprocity thus represents the probability that two nodes in a directed network point to each other. By definition, in an undirected network, reciprocity is always maximum ($r = 1$) because all pairs of nodes are symmetric. The term *dyad*, originally coined by the sociologist Georg Simmel, is commonly used to refer to these pairs of nodes. Formally, dyads are subgraphs in directed graphs consisting of two nodes and the arcs between them.

$$r = \frac{\#mut}{\#mut + \#asym}, \qquad 0 < r < 1, \tag{2.18}$$

where $\#mut$ denotes the number of mutual dyads and $\#asym$ the number of asymmetric dyads.

Taking the definitions of Wasserman and Faust (1994), an *asymmetric dyad* is a pair of nodes that has an arc going in the direction of one node or the other, but not both directions. In turn, a *mutual dyad* is defined by a pair of nodes connected by two arcs, each one going in a different direction (*e.g.*, $a \rightarrow b$ and $b \rightarrow a$, being a and b two nodes in a network).

Density

The density ρ measures the general level of connectivity among nodes in the network, offering clues about, for instance, the speed at which information diffuses among actors. It is defined as the ratio between the number of observed links in the network and the total number of possible links in a network with the same order n, as given in Equation (2.19). Density is a quantity that goes from a minimum of 0, when a network has no edges and all nodes are isolated, to a maximum of 1, when the network is perfectly connected (also termed *complete graph* or *clique*). Therefore, high values of ρ are associated with dense networks, whereas low values of density are associated with sparse networks. In social networks, density gives us insight into, for instance, the speed at which information diffuses among nodes.

$$\rho(G) = \frac{m(G)}{m_{max}(n)}, \qquad 0 < \rho < 1, \tag{2.19}$$

where $m(G)$ is the number of edges that are actually present in the network and $m_{max}(n)$ denotes the maximum number of possible edges in a network also comprised of n nodes. This maximum number is given by $\frac{1}{2}n(n-1)$, when dealing with undirected networks with no self-loops, and by $n(n-1)$ when dealing with directed ones.

Global Clustering Coefficient

The global clustering coefficient C measures the *cliqueness* of local neighbourhoods, thus giving information about the degree to which nodes are tightly bonded in the whole network. High values of this coefficient indicate the presence of cohesive groups of nodes characterised by a high density of links among them, such as *cliques*. Social networks typically display high values of this coefficient due to their transitivity, *i.e.*, the likelihood of finding nodes whose direct neighbours are also connected to each other. The global clustering coefficient is computed as the ratio of closed triplets (or $3\times$ triangles) and the number of triplets (open and closed) in the network, as given in Equation (2.20). Triplets are sets of three nodes that are connected by either two (open triplet) or three undirected links (closed triplet). A triangle comprises three closed triplets, one centred on each one of the nodes. The coefficient is:

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets of nodes}}. \tag{2.20}$$

By definition, $0 \leq C \leq 1$, where 0 indicates a network with no clustering and 1 indicates a network where every node is connected to every other, *i.e.*, everyone knows each other. Opsahl and Panzarasa (2009) proposed a generalisation of this coefficient to weighted networks.

Network Centralisation Index

The network centralisation index C_y proposed by Freeman (1979) measures the degree to which a network is centralised in a small set of actors, by computing the deviations between the maximum centrality score in the network and the centrality scores of the remaining nodes. In order to enable comparison among networks and allow for a better understanding of the meaning of this index, these deviations are normalised by diving their value by the maximum possible deviation for a given centrality measure, thus producing an index that lies between zero and one. The centralisation index of network G with respect to the node centrality measure y (*e.g.*, degree, closeness, or betweenness) is formally defined as follows:

$$C_y(G) = \frac{\sum_{i=1}^{n} [C_y(i^*) - C_y(i)]}{\max \sum_{i=1}^{n} [C_y(i^*) - C_y(i)]}, \qquad 0 \leq C_y \leq 1, \tag{2.21}$$

where $i \in V(G)$, $C_y(i)$ is the score of node i on centrality measure y (*e.g.*, degree, closeness, or betweenness), and $C_y(i^*)$ denotes the largest value of $C_y(i)$ for any node in the network. The denominator of this equation is given by the maximum sum of differences that can be observed for the node centrality measure y. In undirected networks of n nodes, the denominator is given by: (i) $n^2 - 3n + 2$, if y is the degree centrality; (ii) $n^3 - 4n^2 + 5n - 2$, if y is the betweenness centrality; and (iii) $\frac{n^2-3n+2}{2n-3}$, if y is the closeness centrality (Freeman, 1979). Since C_y is a ratio of an observed sum of differences to its maximum value, it ranges from zero to one. The minimum ($C_y(G) = 0$) is

achieved when all nodes have the same centrality scores (*e.g.*, clique networks), and the maximum ($C_y(G) = 1$) is attained if, and only if, there is a single node which globally controls the network due to its central position (*e.g.*, star networks).

In theory, the higher the centralisation index of a social network, the higher the concentration of power in a few number of actors, and the higher the risk of disconnection of the network due to the strong dependence of the network on a few central nodes.

2.5 Link Analysis

In certain types of networks, such as the World Wide Web, one may be interested in finding the most valuable, authoritative or influential node (*e.g.*, web page), or a list of them. To perform this task, several link analysis algorithms were devised, being the Hyperlink-Induced Topic Search (HITS) (Kleinberg, 1999) and the Google PageRank (Brin and Page, 1998; Page et al., 1999) algorithms the most popular ones. These algorithms explore the relationship between hyperlinks to improve the task of information retrieval in the Web, being of extreme importance for the design of efficient web search engines. As the development of these methods was motivated by the problem of web queries, for the sake of simplicity, we will explain them in this context.

2.5.1 HITS Algorithm

Kleinberg (1999) proposed the HITS algorithm to identify the most central web pages for broad search topics in the WWW environment. The specific aim of the HITS algorithm is to find a set of relevant authoritative pages, as well as the set of hub pages that join them together in the hyperlink structure of the WWW pages, in order to retrieve high relevance web pages to a given query (Kleinberg, 1999). In the context of the World Wide Web, a *hub* can be understood as a web page containing a list of hyperlinks to valuable web pages on a given topic. Thus, a hub is a web page that points to many other web pages, *i.e.*, has a high out-degree. The quality of a hub is usually determined by the quality of the web pages (authorities) it points to. On the other hand, *authorities* are web pages that provide reliable contents on a given topic and are cited by many different hubs, *i.e.*, have a high in-degree.

The algorithm explores these two interrelated concepts and estimates the relevance of a web page by computing its authority and hub scores, in an efficient iterative way. After applying the algorithm, the web pages retrieved for a given search query will be those that were ascribed the highest authority score.

Table 2.2: Updated PageRank values after the first iteration $k = 1$.

Node	**A**	**B**	**C**	**D**	**E**	**F**
Shares	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{6}$
Updated PageRank	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

2.5.2 PageRank Algorithm

PageRank is a link analysis algorithm proposed by Brin and Page (1998); Page et al. (1999), which is based on the concept of eigenvector centrality. This algorithm is used by the Google Internet search engine to rank web pages according to the value of the information they carry, in order to ensure that the most valuable web pages appear at the top of the search results.

The intuitive idea behind the algorithm is that information on the World Wide Web can be ranked according to link popularity (the more web pages are linked to a given web page the more popular that web page is). Nevertheless, in this process of weighing web pages, not only the number of links (*i.e.*, the node degree) is relevant, but also the importance of the web pages linking to them. Therefore, PageRank measures the relative importance of a set of web pages based not only on the quantity but especially the quality of their links. In the context of Google's search engine, the obtained PageRank values are used to order the user queries search results so that more relevant web pages are given preference. This way, users have access to higher quality search results.

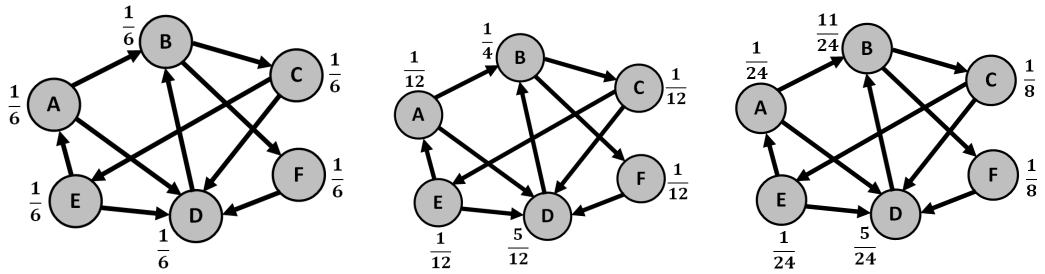
The basic PageRank is computed as follows (according to Easley and Kleinberg (2010)):

Initialisation: In a network of n nodes (or web pages), assign a PageRank value of $\frac{1}{n}$ to each node, and choose the number of iterations k of the algorithm.

1. Update the PageRank value of each node by sequentially applying the following rule:
   Basic PageRank Update Rule: divide the current PageRank value of node p by the number of its outgoing links and pass these equal *shares* to the nodes it points to. Note that if a node p has no outgoing links, the PageRank *share* is passed to itself. The update of a node's PageRank value is performed by summing the shares it receives in each iteration.

2. Apply this rule until the k-th iteration or until convergence.

For the purpose of illustration, consider the following toy example: in a network comprised of six nodes, termed A, B, C, D, E, and F, how can we find the most influential, important or relevant node using the PageRank algorithm? According to the initialisation step of the algorithm, each node is first assigned an equal PageRank of $PR = \frac{1}{n} = \frac{1}{6}$, as represented in Figure 2.2-(a). Then, these values are updated k times (for the sake of brevity, we consider only two iterations) by applying the *basic PageRank update rule*.



(a) Initialisation (b) 1st iteration (c) 2nd iteration

Figure 2.2: Illustration of the process behind the PageRank algorithm in a network comprised of six nodes. The first network (a) corresponds to the initialisation step. In network (b) is shown the updated PageRank values at the end of the first iteration of the algorithm. Note that node D is so far the most authoritative node, with a PageRank value of $\frac{5}{12}$. The rightmost network (c) corresponds to the second (and last) iteration of PageRank. Here, node B overtakes the position of node D in terms of PageRank value.

To apply the rule, first is necessary to compute the shares of all nodes. Then, for each node we sum all shares the node receives. The result of this sum will be its new PageRank value, as shown in Table 2.2 and Figure 2.2-(b). For instance, the share of node D, which has only one outgoing link, is computed as $share(D) = \frac{\frac{1}{6}}{1} = \frac{1}{6}$. Its new PageRank value is given by the sum of the shares of its ingoing links, namely, those coming from nodes A, C, E, and F:

$$PR(D) = \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{6} = \frac{5}{12}. \tag{2.22}$$

After computing these values for all nodes in the network, we repeat the process in the second iteration $k = 2$ and obtain the results reported in Table 2.3 and Figure 2.2-(c). This rule is applied iteratively until convergence of the PageRank values, or until the kth iteration. As we consider only two iterations, we can try to draw some conclusions and interpret the results based only on the information available in Tables 2.2 and 2.3. At the end of the first iteration, node D seemed to be the most promising one, with a PageRank of $\frac{5}{12}$; nevertheless, in the second iteration node B overtakes the position of node D, now occupying the first position in the ranking induced by the PageRank values. This sudden change befits the idea that the PageRank algorithm focuses on the quality, instead of the quantity, of a node's connections. Therefore, and although node D is the one receiving more incoming links, the importance of the nodes linking to them is not as significant as the links of node B. On the other hand, node B has only two incoming links, but these connections carry more value than the connections of node D. Thus, node B gets the highest PageRank value at the end of the second iteration, being considered the most important node in the network. If node B was a social actor, he/she would be considered the most influential one, once this PageRank value means that a great part of the information that flows through the network passes through him/her.

Although PageRank and other link analysis algorithms, such as the HITS algorithm, were

Table 2.3: Updated PageRank values at the end of the second (and last) iteration $k = 2$.

Node	A	B	C	D	E	F
Shares	$\frac{1}{24}$	$\frac{1}{8}$	$\frac{1}{24}$	$\frac{5}{12}$	$\frac{1}{24}$	$\frac{1}{12}$
Updated PageRank	$\frac{1}{24}$	$\frac{11}{24}$	$\frac{1}{8}$	$\frac{5}{24}$	$\frac{1}{24}$	$\frac{1}{8}$

originally motivated by the need to understand the information in the WWW link structure, they are also used in other domains, such as the social sciences. In this field, links can be analysed from two distinct, but somehow interrelated, perspectives: the information-centred and the actor-centred. These perspectives are typically used to help understand the underlying social phenomena, by means of the identification of the most valuable sources of information or, alternatively, the most important actors. Nevertheless, there is still a lack of consensual guiding principles on how to interpret the results of link analysis in a social science context. Thelwall (2006) stresses out the importance of developing guidelines for improving the process of interpreting these results and proposed a theoretical framework for link analysis interpretation.

2.6 Properties of Real-world Networks

In this overview, our focus has been on social networks, which are those resulting from social processes. However, other types of networks can be found in the real world, such as the ones mentioned in Section 2.2 (biological, technological, and information networks). Although these different types of networks correspond to distinct complex systems and distinguish themselves in terms of function or scope, scientists discovered that these networks share a set of common structural properties and exhibit striking regularities that make them unique (Barabási et al., 2009). These findings, which are supported by a wealth of research, offer evidence that the networks we observe in nature and technology converge to similar topologies, and that their evolution is governed by mechanisms that sharply deviate from randomness, contrary to what was earlier believed (Barabási et al., 2009). Before the seminal works of Watts and Strogatz (1998) and Barabási and Albert (1999) on small-world and scale-free networks, respectively, the state-of-the-art model for explaining network formation and growth was the random graph model, which underlied much of the past thinking about real complex systems.

The random graph model was proposed by Rapoport (1953); Erdős and Rényi (1960), and is the simplest and best known model to understand the formation and growth of random graphs. This type of graph is characterised by the random placement of edges between a fixed number n of initially isolated vertices, to create an undirected network in which each of the $\frac{1}{2}n(n-1)$ possible edges is independently present with some uniform probability p. When $p = 0$, we obtain a graph of perfect order - *regular graph*; and when $p = 1$, we obtain a *random graph*, which embodies the chaos.

Since regular and random graphs represent extremes, they are not realistic. Additional properties are required to model complex networks such as the ones we find in the real world. In short, we can say that real-world networks are non-random and non-regular graphs with unique features, where "order coexists with disorder" (Fortunato, 2010, p. 2). As mentioned earlier, a wealth of empirical research has shown that distinct types of real-world networks (*e.g.*, communication, biological, social, and man-made networks) exhibit a broad range of unifying structural properties, such as community structure, presence of motifs, power-law degree distributions, the small-world effect, degree correlations, and large clustering coefficient (Jeong et al., 2000; Camacho et al., 2002; Girvan and Newman, 2002; Sen et al., 2003; Gleiser and Danon, 2003; Andrade Jr et al., 2006; Palla et al., 2007; Leskovec et al., 2008b; Nanavati et al., 2008; Varshney et al., 2011). In this section, we introduce and explain some of these properties for static networks, based on the categorisation proposed by Newman (2003b)[1], which is the following:

1. Small-world effect;
2. Transitivity or clustering;
3. Power-law degree distributions;
4. Network resilience;
5. Mixing patterns;
6. Community structure.

The findings enabled by the empirical observations of a wide range of real-world networks, expressed in these network properties, contributed to a surge of interest in network modelling since the models previously proposed in mathematical graph theory turned out to be inadequate to explain the reality. In light of these new findings, researchers developed new models that reproduce the most significant structural properties observed in real networks. These models allow a better understanding of the mechanisms driving network formation and growth, and can be used to predict the future behaviour of the underlying complex system. Examples of such models include: the small-world network model (Watts and Strogatz, 1998), the preferential attachment model (Barabási and Albert, 1999), the edge copying model (Kleinberg et al., 1999), the forest fire model and the community-guided attachment model (Leskovec et al., 2005, 2007b).

2.6.1 Property 1: the Small-World Effect

Stanley Milgram (1967), an American social psychologist, was the first to formulate the "small-world problem" and to point out the existence of small-world effects in real social networks, through

[1]We only consider six out of the nine properties introduced by Newman (2003b) since we do not cover *degree correlations*, *network navigation* and other structural properties.

a series of famous experiments which are known today as the *Milgram experiment* (Travers and Milgram, 1969). The "small-world problem" was formulated as: *What is the probability that any two people, selected from a large population, will know each other?* (Milgram, 1967). The field experiment conducted to address this problem attempted to test the speculative idea of the small-world effect by tracing acquaintance chains in real social networks that cannot be directly observed, being one of the first direct demonstrations of this effect.

The main hypothesis of the study was that pairs of apparently distant individuals are connected by a short path, *i.e.*, by a few number of acquaintances, through the network. To probe the empirical distribution of the shortest path lengths, in the late 60's Jeffrey Travers and Stanley Milgram arbitrarily asked 296 participants from Omaha, Nebraska, to forward a letter to someone they knew on a first-name basis (*i.e.*, one of their direct neighbours in the underlying social network) in an attempt to get it to an assigned target person living in Boston, Massachusetts (Travers and Milgram, 1969). Out of the 296 letters that were sent, 64 reached the target person by means of a relatively small number of hops (namely, six). With this experiment, it was shown that the median length of the complete chains that succeeded in reaching the target was roughly six, which explains the origins of the concept *six degrees of separation*. Therefore, the main conclusion of the experiment was that we live in a "small-world", since most people are connected by short paths. This effect is a consequence of the high connectivity of real social networks.

The small-world effect is manifested by the existence of shortcuts between most node pairs in a network. In social settings, this means that two apparently disconnected people can quickly get in touch with each other through an incredible low number of acquaintances or friends. This finding has several implications in dynamic processes because it implies, for instance, that the spread of a contagious disease throughout the population will be faster than one would expect. In mathematical terms, the small-world effect means that the average path length between pairs of vertices scales logarithmically, or slower, with the network size with a fixed mean degree centrality (Watts and Strogatz, 1998). This property is also observed in random graphs, where the diameter is very small, growing logarithmically with n, and the vertices have all about the same degree. Watts and Strogatz (1998) developed a network model able to mimic the small-world property found in real-world networks. The networks generated by this model display a large clustering coefficient and a small average path length, which means that any node in the network is connected to any other by a few hops.

The overall conclusion of Milgram has been accepted in a broad sense. In fact, the small-world effect seems to be an enduring pattern and has been widely observed in different types of real-world networks since then (see, for instance, Korte and Milgram (1970); Albert et al. (1999); Dodds et al. (2003)). The advent and establishment of online social networking services over the past decade made possible for social scientists to study very large social structures with unprecedented detail. The once invisible social networks, such as the ones studied by Travers and Milgram (1969), can

now be partially observed based on data made available by individuals through online platforms that map out their social relationships. Although the samples collected through OSN are not controlled random samples, being inherently biased, they provide a fresh view on the characteristics of real-world social networks. An example of such studies is the one conducted by Ugander et al. (2011). These authors analysed the properties of the entire Facebook graph in an anonymised form by focusing on the set of active user accounts reliably corresponding to people. Based on a social graph comprised of 721 million individuals and 68.7 billion friendship edges, they found that the average distance between pairs of Facebook users was 4.7 and that 99.6% of all pairs of Facebook users were within six degrees of separation, which somewhat agrees with the conclusions of the Milgram experiment.

2.6.2 Property 2: Transitivity or Clustering

According to Wasserman and Faust (1994) and Newman and Park (2003), clustering or transitivity is a common property of acquaintance networks, where two individuals with a common friend are likely to know each other. This means that if node i is connected by an edge to node j, and node j is also connected to node z, then there is a high probability that node i is (or will be) connected to node z. This transitive closure in social networks mirrors the natural tendency of people to share common groups of friends and captures the common belief that "a friend of a friend is also a friend".

Mathematically, transitivity is a property of the edges and manifests itself through the existence of many closed triplets of nodes (*i.e.*, three nodes connected to each other by three edges in the network, meaning that every node is connected to the remaining two nodes). Thus, transitivity can be captured by measuring the density of triangles in a network. Transitivity is usually quantified using a clustering coefficient that can be global or local, as mentioned in Section 2.4.

2.6.3 Property 3: Power-Law Degree Distributions

A fundamental quantity commonly measured and reported in the empirical studies of networks is the degree distribution. The degree distribution p_k is the probability distribution of the degrees of nodes over the whole network. Therefore, p_k represents the probability that a vertex chosen uniformly at random has degree k, and is defined by the fraction of nodes in the network that have degree k. This means that, if the total number of nodes in the network is n, and n_k of these nodes have degree k then, for this value of the degree, we have a probability of $p_k = \frac{n_k}{n}$. The probability distribution of the degree centrality in a given network is thus obtained by computing this probability for each degree value k observed in the network.

Random graphs, such as the ones studied by Erdős and Rényi (1960), show a binomial degree distribution because the presence, or absence, of an edge is equiprobable (*i.e.*, equal for all possible vertex pairs). In the limit of large graph size, this binomial degree distribution converges to the

Poisson distribution. Therefore, in this class of graphs the degree distribution is highly homogeneous as most vertices have similar degree. Real-world networks are, in turn, quite different from random graphs with respect to the degree distribution. Barabási and Albert (1999) observed that in real networks the distribution of the node degree is very heterogeneous and highly right-skewed, which means that there is a large number of poorly connected nodes (*i.e.*, with low degree) and a small number of high-degree centrality nodes. Nodes displaying abnormally high degrees are usually called *hubs*. This power-law property has also been observed in the distribution of the number of connected components (Broder et al., 2000) and in the distribution of the size of communities (Toyoda and Kitsuregawa, 2003). The findings of Barabási and Albert (1999) reinforce the results of Price (1965) on networks of citations between scientific papers. In both cases, they found that the degree distribution of several real-world networks, such as citation networks, follow a power-law (at least asymptotically) in their tails: $p_k \sim k^{-\alpha}$, with the α scaling exponent typically varying from $2 \leq \alpha \leq 3$ (Newman, 2005; Clauset et al., 2009). Networks with power-law degree distributions are often referred to as *scale-free networks* because power-laws have the same functional form at all scales (Barabási and Bonabeau, 2003). This kind of distributions draw on the notion of "the-rich-gets-richer" because they usually arise when the amount one gets of something depends on the amount one already has. For instance, in the specific case of degree distributions, the degree centrality can be interpreted as the node's wealth and "rich" nodes have a better chance of receiving new edges than "poor" nodes (*e.g.*, highly cited papers are more likely to be cited again than papers with few or no citations). Price (1976) used the term *cumulative advantage* to refer to this mechanism, which is believed to be the most probable explanation for the power-law degree distributions in several real-world networks, which include, but are not restricted to, citation networks, collaboration networks, and the World Wide Web. Today, this process is best known as *preferential attachment*, a name coined by Barabási and Albert (1999). In their seminal paper, Barabási and Albert (1999) describe a network growth model, which became known as the *Barabási-Albert model* or the *preferential attachment model*, that generates networks with a small number of hubs and a large number of low-degree centrality nodes. Their work shows that a network that grows by preferential attachment will indeed become scale-free, due to "the-rich-gets-richer" strategy employed in the model. Such networks are robust against random node failure but vulnerable to attacks targeted to high-degree centrality nodes. As pointed out by Boccaletti et al. (2006), from an application perspective this knowledge is important as it can affect the efficiency of any process running on top of the network.

2.6.4 Property 4: Network Resilience

Network resilience refers to the impact of removing a fraction of nodes, or edges, on the network connectivity (commonly measured by the diameter or the average geodesic distance), being an indicator of the robustness of the network to random or intentional "attacks". Different kinds of

networks exhibit different levels of resilience. Most networks are robust against random vertex removal (or random edge removal) but considerably less robust to the targeted removal of the highest degree centrality vertices (or of the highest betweenness edges). Also, when the endpoints of a bridge are removed the connectivity of the network is affected, as some of the nodes become disconnected. Betweenness centrality can be used as a measure of resilience as it tells us how many geodesic paths will get longer when a specific vertex is removed from the network. However, in real settings the removal of a single node is not usually cause for alarm, since the networks comprise millions or even billions of nodes. In such cases, it is more appropriate to test the resilience of the network based on the removal of a given fraction of nodes or edges.

The analysis of network resilience has important applications in epidemiology, as it helps the design of effective public health strategies for controlling outbreaks of infectious diseases (Christley et al., 2005). This can be achieved by first identifying the most central individuals in a network of interactions, and then analysing the impact of vaccinating these individuals (equivalent to node removal) in the spread of an infectious disease.

2.6.5 Property 5: Mixing Patterns

Most networks are made up of nodes with different attributes, or characteristics, and the linking between nodes, *i.e.* the probability of connection between two nodes, tends to be selective and dependent on node types (*e.g.*, in a food web representing which species eat which in an ecosystem, nodes may represent herbivores, carnivores, and plants). In social networks, this phenomenon is apparent because, in theory, individuals tend to associate with individuals who are similar to them in one or more characteristics (McPherson et al., 2001). This selective linking is usually called *assortative mixing*, or *homophily*, and three classic examples are mixing by race, gender, and age (Shrum et al., 1988). Observations of real-world social networks agree with this theory, once these networks tend to show positive assortative mixing (Newman, 2003a), *i.e.* people tend to connect preferentially with others who are like them. This tendency may explain some of the well known structural properties of social networks, namely the existence of meaningful community structures, which typically emerge from the preferential interaction among similar groups of individuals. Newman (2003a) also found that assortatively mixed networks are more robust to the targeted removal of their nodes than disassortatively mixed or neutral networks.

Newman (2003a) proposed assortative coefficients to quantify the assortative mixing of a network based on discrete characteristics (*e.g.*, sexual orientation, race, language) or scalar properties of the nodes (*e.g.*, age, income, node degree). These coefficients replace the previously proposed by Gupta et al. (1989) and allow us to distinguish a randomly mixed network from a perfectly assortative one. Newman's coefficients are positive if the network shows assortative mixing, negative if the network displays disassortative mixing (*e.g.*, biological and technological networks), and zero if the network is neutral.

2.6.6 Property 6: Community Structure

Numerous networks found in the real world, such as social and biological networks, display community structure. This means that they organise themselves into sets of densely connected nodes that are weakly connected to other sets of nodes in the network. We delve deeper into this topic in the following section.

2.7 Community Detection

One of the defining characteristics of networks is their community structure. This property has been observed in numerous real-world networks, such as social networks, the World Wide Web, protein interaction networks, and co-authorship networks, and manifests itself as an organisation principle in which networks arrange themselves into cohesive groups, dubbed *communities*, that are locally dense but globally sparse. The recognition of the importance of community discovery for understanding the structural and functional properties of networks, has attracted a great deal of interest across the natural, social, and information sciences (Palla et al., 2005; Boccaletti et al., 2006). This burst of interest in the topic has reflected in a prolific scientific progress, especially in terms of approaches and algorithmic tools for detecting these highly connected mesoscopic structures in networks, turning community detection into a fundamental line of research within network science (Palla et al., 2005; Boccaletti et al., 2006; Leicht and Newman, 2008; Fortunato, 2010).

Communities, which are also known as *modules* or *clusters*, can be straightforwardly defined as groups of entities that are similar in some sense. A widely accepted definition is the one based on edge density and proposed by Girvan and Newman (2002): communities can be understood as natural divisions of the network into densely connected subgroups of nodes, which have few links between them. Thus, we often find high concentrations of edges within certain regions of the graph, called communities, and low concentration of edges between those regions. The high edge density observed in some subsets of nodes is interpreted as a measure of proximity among these nodes. As a result, these communities can be seen as latent concepts that explain many of the observed interactions among entities, such as shared interests, similar buying behaviour, same social roles, and same occupations. Note that the above definition assumes that communities are defined by their structure, *i.e.*, by the arrangement of edges that connects their members. A closely related definition is the classical one, which sees communities as cliques, *i.e.*, as subgroups of nodes that are all directly connected to each other. Since for several applications this is an overly conservative and strict definition of community, less restrictive definitions were proposed over the years, such as s-clique, s-club, s-plex, k-core, and k-block (Pattillo et al., 2013). Alternative definitions of community exist and are often based on the similarity of nodes' attributes or semantics. These draw on the logic of classical data clustering (Jain, 2010) and typically neglect the relationships

established among the nodes.

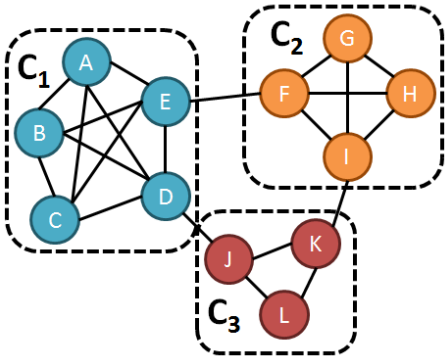


Figure 2.3: Illustration of a network with three distinct communities: $C_1 = \{A,B,C,D,E\}$, $C_2 = \{F,G,H,I\}$ and $C_3 = \{J,K,L\}$.

To better understand the concept of community, Figure 2.3 depicts a simple network comprising three communities, named C_1, C_2, and C_3. In this picture, we represent an ideal scenario since each community is itself a complete graph, or a clique, of varying size ($C_1 = K_5$, $C_2 = K_4$, and $C_3 = K_3$). Also, the density of ties between communities is very low. The few ties that exist are bridges, since they are the only available connections between different regions of the network. In real life, we can find several examples of such tight groups. Society is a rich environment for finding communities, once people have the natural tendency to form groups. These groups can be families, circles of friends, work-related groups, religious groups, towns, nations, and so on. If we also consider groups formed by companies or by customers, we can identify communities with relevance to the business field, such as communities of companies working together in the supply chain (*e.g.*, manufacturer, wholesaler, retailer, logistics operator) or communities of customers with similar purchasing habits. Biology is another field where methods for finding communities are important, especially within the scope of metabolic networks (Jeong et al., 2000) and protein interaction networks (Rual et al., 2005). For instance, the analysis of the topology of protein interaction networks allows to uncover functional building blocks of the network, *i.e.*, groups of proteins with similar functions within the cell. We can also find virtual online communities in the World Wide Web, or groups of topic-related web pages, which may be useful for the development of automatic and efficient recommendation systems.

Although the concept of community is rooted in a structural characteristic (*e.g.*, a high edge density), the internal structure of communities is not necessarily similar in terms of density, connectivity, and centrality. For this reason, after detecting communities in the network it may be interesting to analyse their topologies by computing classical node-level and network-level measures. For instance, the analysis of the structural position of nodes in centralised communities can help identify central and intermediate actors. Central actors are those occupying central

positions in the community they belong to, being often associated with group control and stability functions. On the other hand, intermediate actors are those who lie at the boundaries of the community, playing a key role in the spread and exchange of new ideas and information due to their ability to create bridges between communities. This knowledge not only contributes to a better understanding of the underlying system but it can also be harnessed for specific applications, such as selecting a suitable target for a given Marketing campaign or vaccination program.

According to Newman (2006), there are two strands of research on discovering communities in network data. The first has its origins in computer science and is known as *graph partitioning*, whereas the second has been mainly pursued by sociologists (and more recently, by physicists, biologists, mathematicians, and data scientists) being usually referred to as blockmodelling, hierarchical clustering, or community structure detection. The former arose in the computer science field and was motivated by the need to find the best way to allocate tasks to processors so as to minimize the communications between them. This network optimisation task aimed at enhancing the computation in parallel computing environments by splitting the computer cluster into groups containing approximately the same number of processors. One of the earliest methods proposed for solving the graph partitioning problem was the *Kernighan-Lin algorithm* (Kernighan and Lin, 1970). The latter was motivated by the observation of community groups within society and its goal was to simplify the analysis of social phenomena through the arrangement of people according to their similarities. The graph partitioning problem differs from the community structure detection in several aspects: the size of the communities are usually known in advance and the goal is to find the best partition of the graph even if a good division does not exist. The methods and algorithms that will be introduced in this overview fall within the second strand of research.

The following subsections are devoted to the introduction of some classical methods for community discovery. The majority of these methods assume partitions of nodes, *i.e.*, they find disjoint communities, so each node is assumed to belong to a single community. However, many real-world networks are made of highly overlapping cohesive groups of nodes. For instance, people play several roles in society, simultaneously belonging to several communities (*e.g.*, family, work, friends, school, sports), which makes this approach appealing for the social sciences. When dealing with networks whose nature implies the existence of overlapping communities, it is desirable to use methods specifically tailored to this problem. Several methods were proposed to this aim, such as the ones by Baumes et al. (2005), Palla et al. (2005) and Derényi et al. (2005), Gregory (2007, 2010), Evans and Lambiotte (2009), and Nicosia et al. (2009). The most popular of these methods is the Clique Percolation Method (CPM) proposed by Palla et al. (2005) and Derényi et al. (2005). The idea behind CPM is that edges inside a community are likely to form cliques due to their expected high density. Cliques are fully connected subgraphs in the network and their size is usually denoted as k. In the CPM a community is, thus, defined as the maximal union of k-cliques that can all be reached from each other via a sequence of adjacent k-cliques. By definition, these

k-clique communities can naturally share nodes, which allows the CPM to uncover the overlapping community structure of a network. Although it works well in practice, the CPM method relies on a strict condition to find groups in the network, since it requires that every node is connected to any other inside the community, which may be unrealistic in certain contexts. A more relaxed concept is that of *quasi-clique* since it only requires a certain proportion of the links, instead of all links, to be present among the nodes.

2.7.1 Hierarchical Clustering

Hierarchical clustering is a popular class of methods for finding clusters, since it does not require any assumptions regarding their number, membership, and size. Hierarchical clustering algorithms produce a flexible nested structure (smaller clusters within larger clusters which, in turn, are embedded in even larger clusters), typically represented by means of a dendrogram, that uncovers the multilevel structure of the network. These characteristics are pertinent for problems where little information is available concerning the community structure of a network, or for problems where it is crucial to understand the various levels of organisation of the nodes in the network. In addition, hierarchical methods proved to be quite effective in solving cluster analysis problems, thus becoming attractive for graph partitioning and community detection purposes.

The procedure behind traditional hierarchical clustering is quite intuitive since it is based on the definition of similarity. Usually, the first step is to select the similarity measure that will be used to assess how alike two nodes are according to a given global, or local, property. Examples of such measures are the cosine similarity, the Jaccard index, the Euclidean or Manhattan distances, the Hamming distance between pairs of rows in an adjacency matrix, among others. The next step is to compute the similarity matrix between all pairs of nodes, regardless of the fact that those nodes are, or not, connected to each other. Then, one chooses the approach to group them - the agglomerative or the divisive - and, depending on the choice, selects a given distance measure to compute the similarity between clusters (*e.g.*, single linkage, complete linkage, Ward's method, etc.). The result is a dendrogram depicting the arrangement of clusters returned by the hierarchical algorithm. Since clusters often display an internal clustering structure, *i.e.*, they comprise smaller clusters, which in turn also contain even smaller clusters, the dendrogram provides a natural way to represent the hierarchical structure of the graph at each possible granularity level. However, for problems requiring a single network partition, one needs to select at which level the dendrogram should be cut. This is equivalent to choosing the number of communities k. A typical strategy is to compute the value of *modularity* (Newman and Girvan, 2004) for every possible number of communities, or dendrogram level, and select the number that maximizes this function.

As mentioned before, there are two general approaches for hierarchical clustering, which are the following:

Divisive methods: this class of methods focuses on detecting and removing the links that connect nodes belonging to different regions in the network, so that densely connected regions (*i.e.*, communities) get disconnected from each other. These links are quite often bridges or local bridges between network regions. The general procedure of divisive methods is to start by considering the whole network as a single community and then perform successive divisions of this large community into smaller communities. A well known algorithm exploring this method for community detection is the one proposed in the seminal paper of Girvan and Newman (2002).

Agglomerative methods: in contrast with divisive methods, this class of methods focuses on the tightly knit parts of the network rather than on the connections at their boundaries. Methods belonging to this class start by considering each node as a community and, at each step, combine these communities into more inclusive communities until only one community remains. Walktrap (Pons and Latapy, 2005) and the Louvain method (Blondel et al., 2008) are two examples of hierarchical agglomerative methods for community detection.

In the next subsection, we present the most popular divisive hierarchical algorithm for finding communities: the algorithm of Girvan and Newman (2002). We also introduce the Louvain method, which is an agglomerative hierarchical method, in the subsection devoted to modularity optimisation. For a more comprehensive description and discussion of the community detection problem and corresponding methods, we point the interested reader to the excellent survey of Fortunato (2010); Fortunato and Castellano (2012).

Girvan-Newman Algorithm

The most popular algorithm for community detection is the one devised by Girvan and Newman (2002) and known as the *Girvan-Newman algorithm*. This algorithm is also historically important because it laid the groundwork for the modern research in community detection in graphs.

The Girvan-Newman algorithm is a hierarchical divisive technique that divides the original network into disjoint communities based only on the information provided by the network topology. The original network is deconstructed into progressively smaller connected pieces, until the point where there are no edges to remove and each node represents itself a community. Based on the notion of communities as densely connected groups of nodes that are sparsely connected between them, the authors propose the edge betweenness centrality b_e as a measure to identify community boundaries. Edges with high betweenness centrality lie on a large number of shortest paths between nodes and, therefore, are believed to connect different non-overlapping communities. Hence, the main idea of the algorithm is to isolate the existing communities in a network by means of the identification and iterative removal of inter-community edges.

The input of the algorithm is a graph and the output is a hierarchical structure or dendrogram, where communities at any level correspond to a horizontal cut through the dendrogram at a given height. The steps of the algorithm can be summarised as follows:

1. Compute the betweenness centrality of all edges in the network;
2. Remove the edge with highest betweenness centrality b_e. This step may cause the network to split into separate disconnected parts. These parts constitute the first level of the dendrogram;
3. Recompute the edge betweenness centrality in the running network;
4. Repeat the previous steps until there are no edges to remove in the graph.

The original version of the Girvan-Newman algorithm did not include any procedure to select the best network partition. However, in a further refinement of the algorithm (Newman and Girvan, 2004), the modularity function has been used to support this choice. The partition associated with the largest modularity value is assumed to be the best one. The modularity function has been frequently used ever since to evaluate the quality of community structures.

Because of its popularity, this algorithm is integrated in several libraries of well known network analysis programs. For instance, in `R` (Team, 2008) one can use the `edge.betweenness.community` function, from the `igraph` library, to apply the Girvan-Newman algorithm.

Despite its simplicity, the Girvan-Newman algorithm is only suitable for networks of moderate order (up to a few thousand nodes), due to the high computational cost involved in the computation of the edge betweenness centrality at some steps of the algorithm. Since the edge betweenness centrality relies on the computation of all pairs shortest paths, its values depend on the properties of the whole network, thus being a global measure. The computational cost of the Girvan-Newman algorithm hampers its application to large networks, such as the ones available today. To overcome this problem, Radicchi et al. (2004) proposed a similar hierarchical divisive community detection method that improves the computational efficiency of the Girvan-Newman algorithm, by introducing a local measure, the edge clustering coefficient, as the criterion to remove the inter-community edges. By using a local measure (edge clustering coefficient) instead of a global measure (edge betweenness centrality), this method significantly improved the computational complexity of the Girvan-Newman algorithm, from $O(N^3)$ to $O(N^2)$ on a sparse graph.

2.7.2 Modularity Optimisation

A widely used class of methods to detect communities in networks is modularity optimisation. Modularity Q is a quality function, proposed by Newman and Girvan (2004), that evaluates the merit of a given partition of the network into communities. This measure is useful to judge the

goodness of the communities returned by a given community detection algorithm in real-world networks, since for numerous problems there is no *a priori* knowledge of the true community structure of the network and, thus, no way to quantitatively measure the quality of the results. Modularity has been used not only to compare the quality of the partitions obtained by different community detection methods, but also as an objective function to optimise. According to Newman (2006, p.8578),

> Modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random.

The idea underlying this measure is that a good partition of the network into communities is the one in which the number of edges between (within) communities is less (larger) than what would be expected on the basis of random chance. In other words, it is reasonable to assume that if the number of edges between (within) communities is significantly less (larger) than one would expect by chance, there is evidence of a meaningful community structure in the network. The larger the deviation between the real and the expected number of edges, the more *modular* the network is. The idea behind modularity requires the definition of a null model, *i.e*, a comparison model that creates a structurally identical copy of the network under analysis but without community structure. Usually, the random graph model of Erdős and Rényi (1960) is used for this purpose because, since any pair of nodes have the same probability of being connected, there is no preferential linking mechanism that would generate a community structure. In order to perform fair comparisons, the network of the null model should match some of the structural properties of the network being analysed (for instance, the average degree centrality and the number of nodes should be the same). In its standard formulation, modularity measures the degree to which the network being analysed deviates from this null model. The larger the deviation, the stronger the evidence of the meaningfulness of the community structure present in the analysed network. Thus, large positive modularity values are, by assumption, associated with good network partitions since they indicate that the found communities are internally denser and externally sparser than would be expected in a random graph with similar structural properties. This assumption motivated the development of modularity optimisation methods for community detection, which attempt to find the best network partition, or at least a very good one, by maximizing Q.

Modularity is computed as:

$$Q = \frac{1}{2m} \sum_{ij} \left[a_{ij} - \frac{k_i k_j}{2m} \right] \delta(Com^i, Com^j), \tag{2.23}$$

where m is the number of edges, k_i and k_j represent the degree centralities of nodes i and j, respectively, a_{ij} is the entry of the adjacency matrix that indicates if nodes i and j are connected or not, $\frac{k_i k_j}{2m}$ represents the expected fraction of edges falling between nodes i and j in a random graph with a relatively homogeneous degree distribution, Com^i and Com^j denote the groups to which nodes i

and j belong, and $\delta(Com^i, Com^j)$ represents the Kronecker delta, which is 1 if nodes i and j are in the same community and 0 otherwise. The sum $\sum_{ij}$ means that the deviation is computed over all pairs of nodes in the network.

Modularity Q can be either positive or negative, and it is always smaller than one (Fortunato and Castellano, 2012). Large negative values signal the existence of groups with few internal edges and many inter-group edges, which is basically the opposite of a modular network. Partitions in which each node is a community also yield negative values of modularity. Hence, a negative Q is a good indicator of the absence of community structure in the network. In turn, positive values reveal the possible presence of community structure in the network. If Q is not only positive, but also large, then the corresponding partition may reflect the natural division of the network into communities. According to Clauset et al. (2004), it was empirically found that a modularity equal or larger than $Q \geq 0.3$ is a good indicator of the existence of meaningful communities. Throughout this document, namely in the next two chapters, we use the expression *modularity of a network* to refer to the largest modularity value returned by a given community detection algorithm, and we also assume that a network partition is meaningful if $Q \geq 0.3$.

Assuming that the larger the modularity the better the obtained network division is, a natural approach for tackling the community detection problem is to maximize this function. In theory, this can be done by computing the modularity for every possible partition of the network and then selecting the partition returning the highest Q value. This simple idea inspired a new class of methods whose foundations are set on the maximization of modularity. Albeit this approach is an effective way to address the problem, the exhaustive search over all possible network divisions is computationally intractable since the general problem of modularity optimisation is NP-complete and cannot be solved in polynomial time (Brandes et al., 2007). This problem has been circumvented by adapting a number of heuristic methods to this specific optimisation problem. Using heuristics, it is possible to obtain a fairly good approximation of the global optimum (in this case, the maximum value of modularity) in an acceptable time. Algorithms that employ this strategy are, for instance, the one proposed by Blondel et al. (2008), which performs a hierarchical optimisation of modularity by exploring greedy techniques, and the one proposed by Guimera and Amaral (2005), that applies the simulated annealing procedure to the modularity optimisation problem. In this subsection, we only introduce the Louvain method because, in addition to being considered one of the best available methods for community detection (Lancichinetti and Fortunato, 2009), is the one we use to find communities in networks in chapters 3 and 4.

Louvain Method

The Louvain method is a heuristic method based on a greedy hierarchical optimisation of the modularity function Q. It belongs to the class of hierarchical agglomerative methods since it

recursively merges similar nodes/communities, in a bottom-up fashion, to produce a hierarchy that reveals the structure of the network at different levels of organisation. It is also a modularity optimisation method since it performs a local optimisation of the modularity function. In its original formulation, the Louvain method uncovers the hierarchical structure of the network by finding several levels of partitions. Each partition is generated at each *pass* of the algorithm. Each *pass* comprises *two phases*: (i) the local modularity optimisation phase and (ii) the community aggregation phase. The height of the final hierarchy is determined by the number of passes of the algorithm.

The Louvain algorithm starts by considering each node i ($i = 1, ..., n$) as a distinct community, which means that the first partition comprises n communities. Then, for each node/community i, it looks for positive *gains in modularity* by moving the isolated node from its community to a neighbouring community (*i.e.*, a community to which node i is connected). Node/community i is then merged with the community associated with the maximum positive modularity gain. In case no positive gains exist, the node/community i stays in its community. This search for a local optima is repeatedly and sequentially applied for all nodes/communities in the network until no further improvements are possible. According to the authors, the gain in modularity ΔQ obtained by moving an isolated node i to a neighbouring community Com is given by the following equation:

$$\Delta Q = \left[\frac{\sum_{in} + k^w_{i,in}}{2m^w} - \left(\frac{\sum_{tot} + k^w_i}{2m^w} \right)^2 \right] - \left[\frac{\sum_{in}}{2m^w} - \left(\frac{\sum_{tot}}{2m^w} \right)^2 - \left(\frac{k^w_i}{2m^w} \right)^2 \right], \tag{2.24}$$

where $\sum_{in}$ denotes the sum of the weights of the edges inside community Com, $\sum_{tot}$ gives the sum of the weights of all edges incident on nodes in Com, k^w_i is the strength or weighted degree centrality of node i, $k^w_{i,in}$ gives the sum of the weights of the edges connecting node i to nodes belonging to Com, and m^w denotes the sum of the weights of all edges in the network.

At the end of this phase, a partition of the network is identified. In the second phase, the previous network is converted into a much smaller meta-network, where nodes are replaced by super-nodes. These super-nodes represent the communities detected in the first phase (each super-node is, therefore, an aggregation of nodes). Links between super-nodes are weighted by the sum of the strength of the edges connecting nodes that belong to the corresponding two communities. On the other hand, links connecting nodes inside the same community are represented by self-loops in the meta-network. This two-phase procedure is then iterated until a maximum of modularity is attained, and new hierarchical levels and meta-networks are yielded. The algorithm stops when modularity converges to a value where no more gains are possible.

The Louvain algorithm is sensitive to noise, non-deterministic, and order-sensitive, as it might return different outputs depending on the order in which the vertices are considered in the sequential analysis of modularity gains. Nonetheless, it is able to extract high-quality partitions from large networks in a fast way and with low computational cost, since its computational complexity is

essentially linear with the size m of the network. Besides, it is parameter-free, highly intuitive, and easy to implement. Based on the study of Lancichinetti and Fortunato (2009), who performed an extensive comparison of several community detection algorithms, the Louvain method was considered one of the best available methods, achieving an excellent performance with the additional advantage of low computational complexity.

2.8 Summary and Conclusions

Professor Stephen Hawking envisioned the 21st century as the "century of complexity". This remark befits the idea that our world is becoming even more complex over time and there is an urgent need to understand complexity by adopting a systems thinking. Inherited from the realms of physics and mathematics, systems thinking is grounded on the observation of reality as a whole, as an alternative to the traditional perspective of looking at reality as a set of isolated elements, thus bringing relationships and interactions to the forefront of the research focus. Despite the remote origins of Social Network Analysis, which aimed at understanding complex social structures through the network perspective, recent technological advances in computational power and storage capabilities made it possible to collect and store huge amounts of network data, naturally opening up exciting opportunities for network research and the study of complex systems. The explosive growth in data coupled with modern computational resources enabled important research discoveries on how complex biological, technological, and social networks organise, operate, and behave, leading to important insights in a diversity of fields. Some of the most important principles and non-trivial properties of complex networks discovered in the last years, such as the power-law degree distribution and the community structure, were briefly introduced in this overview, as well as the main concepts, definitions, tasks, and methods of network research.

Since the inception of SNA, most network research has focused on the study of static networks, thus neglecting their dynamic nature. However, complex networks are not static but rather evolve over time. Therefore, the incorporation of temporal characteristics in network analysis is a key source of interest, contributing to a deeper understanding of how complex systems emerge, grow, and change across time. Several studies already explored this research direction and the last decade witnessed a growth of attention towards the modelling and characterisation of dynamic complex networks (Boccaletti et al., 2006; Barrat et al., 2008; Aggarwal and Subbian, 2014). The interdisciplinary nature of network research and the examination of the multiple facets of complex networks, are paramount for achieving the ambitious objective of arriving at a unifying picture of complex networks that lays the groundwork for developing the long-sought general theory of complexity.

In the next chapters, we will draw on the importance of understanding the dynamics of networks by introducing methodologies for analysing and visualising evolving networks. These

methodologies are intended to help organisations better understand their market and customers by shedding light on their temporal changes.

Chapter 3

Dynamic Communities in Evolving Customer Networks

The widespread availability of CRM applications in modern organisations, allowed companies to collect and store vast amounts of highly detailed customer-related data. Making sense of these data using appropriate methods can yield insights into customers' behaviour and preferences. The extracted knowledge can then be explored for marketing purposes. SNA techniques can play a key role in business analytics. By modelling the implicit relationships among customers as a social network, it is possible to understand how patterns in these relationships may translate into competitive advantages for the company. Additionally, the incorporation of the temporal dimension in such analysis can help detect market trends and changes in customers' purchasing behaviour and preferences. In this chapter, we introduce a methodology to examine the dynamics of customer communities, or profiles, which relies on two different time window models: a landmark and a sliding window. Landmark windows keep all the historical data, and treat all nodes and links equally, even if they only appear at the early stages of the network life. Such approach is appropriate for the long term analysis of networks, but may fail to provide a faithful picture of the current evolution. On the other hand, sliding windows focus on the most recent past thus allowing to capture better the recent changes of the network. The application of the proposed methodology on a real-world customer network suggests that both window models provide complementary information. Thus, in cases where there is no information regarding the dynamicity of the customer network, we recommend using together these two window models for the purpose of temporal analysis of customer communities.

3.1 Introduction

The scientific and technological advances of the last decades permeated virtually every facet of our everyday lives, revolutionising the way how people interact, communicate, work, buy, and access information. These advances also shaped the market and, as a consequence, how business organisations operate and relate with their customers. The proliferation of competitors and the weakening effectiveness of traditional marketing promotional campaigns (Leskovec et al., 2007a; Lee, 2012) had led companies to evolve from product-centred strategies to customer-centred strategies (Miguéis et al., 2012). Besides, customers of the information age are becoming more informed and demanding, and their needs tend to change frequently. The characteristics of this "new" customer, coupled with a constantly evolving market, pose additional challenges to the prosperity of companies. In face of these challenges, companies willing to thrive in the market need to use their best endeavours to learn about their customers' needs in order to respond or even anticipate changes in these needs, in a timely and accurately manner. An effective way to achieve this goal is to leverage customer data and analytics.

The widespread availability of CRM applications in modern organisations, allowed companies to collect and store vast amounts of highly detailed customer-related data (*e.g.*, purchasing habits, values of proposals, demographic variables). Making sense of these data using appropriate methods can yield insights into customers' behaviour, needs, and preferences. The extracted knowledge can then be used to support the redesign of marketing promotions or to set up marketing campaigns tailored to each individual customer, or to a group of customers showing similar purchasing behaviour or preferences. This kind of analysis can be performed using SNA techniques, by modelling the implicit relationships among customers as a social network. We define *customer network* as a finite set of customers who are linked to each other if they bought the same product during a given time frame. These links are weighted by the number of common products purchased by pairs of customers. Modelling customer-related data using this customer network model has the advantage of uncovering implicit relationships among customers based on the similarity of their purchasing behaviour over a given time period. Furthermore, since purchasing events are annotated with timestamps, it is possible to extract the network state at different moments in time, thus enabling the study of customer network evolution. Additionally, if we analyse these dynamics at the community-level, we are able to identify evolutionary profiles of groups of customers that exhibit similar purchasing behaviour. The identification of customer profiles based on customer purchasing behaviour, and the analysis of their temporal development, enables the discovery of patterns of customer needs and an in-depth understanding of how these needs change over time. Businesses can then use this knowledge to nurture their customers by responding more effectively to their needs or by anticipating their future needs. We assume that the explicit purchasing behaviour of customers reveals their needs. In this chapter, we propose a methodology to both discover and track customer

profiles over time by employing SNA techniques and community evolution analysis methods. We also explore two time window models to capture changes in these profiles at different levels of temporal resolution. From the application point of view, to the best of our knowledge the most related research to ours is the one by Böttcher et al. (2009). Their work proposes a methodology for identifying and monitoring customer segments over time, but their approach is methodologically different since it is based on the discovery of frequent item sets. Next we motivate the use of SNA for addressing this problem.

Albeit the origins of network studies go back a few decades ago, in recent years we witnessed an impressive advance in network-related fields, especially in computer science and computational physics. Until recently the analysis of such networks was mainly a static investigation of the aggregated graph of the network across multiple snapshots, which discarded information regarding the temporal dimension (Berger-Wolf and Saia, 2006; Tantipathananandh et al., 2007; Takaffoli et al., 2011). Nonetheless, one of the key features of many real-world networks is that their topology is not static but rather evolves over time, as new nodes and links are added to the network, old ones disappear, and the interaction among nodes changes with time. Therefore, approaches focusing on the analysis of a fixed or aggregated snapshot of the network may fail to capture the dynamic behaviour of the evolving network and may obscure the temporal patterns. As a result, decisions based solely on static analysis of dynamic networks are potentially flawed and inaccurate. Harnessing the richness of information contained in a dynamic network is therefore relevant and useful as it helps understand how the network has evolved, *i.e.*, the stages it had to go through before it reached the current state, and provide clues about future states or changes.

Since the formation and changes undergone by communities reflect the dynamics at the whole network, frameworks to model and track the life-cycle of communities within dynamic networks have been developed (Toyoda and Kitsuregawa, 2003; Hopcroft et al., 2004; Berger-Wolf and Saia, 2006; Falkowski et al., 2006; Palla et al., 2007; Asur et al., 2009; Greene et al., 2010; Takaffoli et al., 2011; Bródka et al., 2013). The majority of these frameworks belong to a thread of research that models dynamic networks as temporally ordered sequences of network snapshots and whose analysis relies on a set of community structures independently learned for each network snapshot. This snapshot-based approach implicitly assumes that time is discrete and has the advantage of assuming non-stationary data generating processes (Spiliopoulou, 2011), thus being more general and realistic. These community structures are detected at different stages of the evolving network, by applying a suitable community detection algorithm to each snapshot of the network. These snapshots can be extracted by considering the whole data accumulated until the corresponding time point, or by considering the data that has been observed during a given time interval. The former corresponds to a *landmark window* strategy and the latter to a *sliding window* strategy. The landmark window model typically assigns the same weight to nodes and links, even if they were only active at a remote time point, and is suitable for detecting persistent or enduring patterns in

the evolving network, thus providing an overall view of the evolution trends. In turn, the sliding window model assigns different weights to more recent observations, thus focusing on the most recent state of the dynamic network when analysing its evolution. Hence, this window model is able to capture more up-to-date events, especially when dealing with volatile and highly dynamic networks.

Despite the burgeoning literature addressing the problem of dynamic community analysis, to the best of our knowledge few work to date has explicitly explored the effect of selecting different time window models (also referred to as *timeframe types* in the literature) on both the stability of dynamic communities and the type of information provided by each model. In this context, the most related work to ours is the one by Saganowski et al. (2012). These authors carry out an empirical study of the influence of several time window models (*e.g.*, sliding window with no overlap, overlapping sliding window, landmark window) on the number of detected community events (*e.g.*, forming, shrinking, merging, splitting). Based on their experiments, they conclude that the choice of the granularity at which the time-varying network is snapshotted impacts the results of the method used to extract community dynamics. While landmark windows prove to be useful to detect persistent, enduring, and stable communities, the sliding window model with overlap is more suitable for extracting community evolution in rapidly changing social networks. On the other hand, extracting network snapshots for disjoint time frames precludes the creation of complete evolutionary profiles of communities, since the detected changes are too fast (*e.g.*, formations followed by dissolutions). In our case study, we draw similar conclusions. Another related research is the one by Falkowski et al. (2006), who developed a two-pronged framework to analyse the evolution of two types of dynamic communities in social networks: communities with rather stable membership structure and communities with high fluctuation of members. For both scenarios, they use an overlapping sliding window approach to obtain the snapshots of the underlying network. On the other hand, the work by Greene et al. (2010) on the same topic suggests that the size of the time step window can influence the obtained results, especially if the network structure is unstable. Mei and Zhai (2005) discuss the impact of the choice of the size of the time window length in the analysis of evolution, stating that a smaller window length allows to detect local temporal patterns whereas a larger window length allows to detect overall evolution patterns. Kawadia and Sreenivasan (2012) also stress out the importance of determining the granularity of the temporal snapshots for the purpose of detecting temporal communities and argue that there is a natural multiscale of interest, driven by the application, for generating these snapshots. However, none of these works compared distinct time window models within the scope of a real-world marketing application.

To fill this gap, in this chapter we propose an application-driven methodology to study evolving customer networks at different levels of temporal resolution. The introduced methodology allows to discover and monitor customer profiles over time, and to detect different types of changes, or

events, in their evolution. Our contributions are threefold: (i) extension of a previously proposed event-based framework for monitoring clusters dynamics, dubbed Monitor of the Evolution of Clusters (MEC) (Oliveira and Gama, 2012), to tackle the problem of community evolution; (ii) application of community evolution analysis, using the proposed event-based framework, in a real-world customer network, with the purpose of identifying and tracking different evolutionary profiles of customers, and (iii) comparison of the type of information delivered by two different time window models, the landmark window and the overlapping sliding window, in the analysis of the evolution of customer profiles.

This chapter proceeds as follows. Section 3.2 provides the necessary background on social network analysis, community detection, and window models. Section 3.3 discusses related work on dynamic network analysis. In Section 3.4, we outline the proposed methodology and provide a detailed description of the extended MEC, termed MECnet. Our case study on a real-world evolving customer network is presented in Section 3.5. Section 3.6 concludes the chapter and discusses directions for further research.

3.2 Background

3.2.1 Social Network Analysis

Social network analysis is a quantitative methodology whose development significantly benefited from the collaborative efforts of academic researchers from different scientific areas (*e.g.*, social psychology, sociology, physics, biology, computer science). SNA offers a powerful means to model, describe, and analyse network structures, groups of nodes (*i.e.*, communities), and single nodes by focusing explicitly on the relationships, or interactions, established between them (Wasserman and Faust, 1994). The focus on the relationships rather on the entities themselves is a fundamental axiom in social network analysis. This axiom stresses the notion that nodes are not independent but rather influence each other. Important lines of research in social network analysis and network science involve the identification of the most prominent nodes, the estimation of their roles within the overall network structure, community detection, link prediction, and the discovery of persistent patterns of relationships and emergent properties that help explain network formation and growth. Several SNA measures were proposed to assess the overall structure of social networks and to measure the centrality of single nodes. The former encompasses measures such as density, clustering coefficient, diameter, average geodesic distance, and average degree. The latter includes measures such as degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and local clustering. We will make use of these measures to characterise the customer network topology at different stages of its evolution.

3.2.2 Community Detection

One of the unique topological features of real-world social networks is their *community structure* (Newman, 2003b). These mesoscopic structures give insight into the high-level organisation of entities within the network, providing a structural summary of the network that allows to uncover existing behavioural patterns and improve the understanding of the underlying complex system. Community structure in a network usually arises as a consequence of both global and local heterogeneity of links' distribution in a graph. Thus, we often find in networks tightly connected groups of nodes, termed *communities*, which are sparsely connected to other densely connected groups. The identification of these unusually cohesive groups often reveals interesting properties shared by the members, such as common social roles, hobbies, interests, values and viewpoints, and professional occupations. The task of *community detection*, which aims at finding meaningful group structures in networks, is itself an important and intensively studied strand of research on the field of SNA and a significant number of methods and algorithms have been proposed for this purpose (for a thorough review, please refer to Fortunato (2010); Fortunato and Castellano (2012) and Xie et al. (2013)). In this chapter, we resort to the Louvain method (Blondel et al., 2008) to detect meaningful communities at each snapshot of the evolving network, although we also perform experiments with the Label Propagation (LP) algorithm (Raghavan et al., 2007). These algorithms belong to the broad category of static community detection methods.

The Louvain method is a greedy optimisation method that uncovers the hierarchical structure of a network by recursively merging communities in a way that improves the *modularity* of the network (Newman and Girvan, 2004). This method generates a hierarchy of partitions, whose height is determined by the number of passes of the algorithm. Each pass of the algorithm comprises two phases. The first phase optimises modularity in a local way by looking for positive gains in modularity when moving a node to a neighbouring community. In the second phase, a new network is built. In this network, nodes pertaining to each one of the communities detected during the previous phase are aggregated into super-nodes. This two-phase procedure is iterated until modularity converges to a value where no more gains are possible. This method is parameter-free and produces very good quality partitions in a very fast way.

The LP algorithm is another commonly used method for extracting communities from networks that relies on the network structure alone to find densely connected groups of nodes. The basic idea behind LP is to explore the information diffusion enabled by the network structure to identify consensual groups of nodes' labels. It starts by labelling every node with unique labels. Then, through an iterative process, the labels are updated by majority voting on the neighbourhood of the node. When this process stops, the communities will correspond to sets of nodes sharing the same label. LP has similarities with the Louvain method in the sense that it is computationally efficient and does not require any *a priori* information about the communities (*e.g.*, number or size of communities, central nodes) to operate. However, while the Louvain method is modularity-based,

relying on a two-stage hierarchical modularity optimisation to detect communities, the LP algorithm does not require the optimisation of a pre-defined objective function to identify network partitions.

3.2.3 Window Models

Landmark Windows

Landmark windows (Gehrke et al., 2001) encompass all the data from a specific point in time up to the current moment. This model is initialised by first selecting a fixed time point (the so-called *landmark*), which marks the beginning of the time window, and then it grows the window by considering all the data seen so far after the landmark. In the dynamic social networks setting, a landmark window will aggregate the network data (*e.g.*, nodes and links) observed over the entire period of observation. By keeping track of all the connections and nodes in the network, this approach does not entail loss of information. However, since it relies on the accumulation of data over time, it is not very well suited to find current trends. Thus, recent interesting phenomena may go unnoticed due to the smoothing effect on data changes occurring over time.

Sliding Windows

Unlike landmark windows, the sliding window model (Datar et al., 2002) incorporates a time-based forgetting mechanism, keeping only the latest information inside the window and disregarding all the data falling outside the window. The simplest approach are sliding windows of fixed length. The window length is a user-defined parameter which influences the amount of data taken into consideration in the model. The *time-based length* sets the window length as a fixed time span. By deeming only the most recent past, this model proves useful in finding current trends. As a drawback, it might be hard to determine the right parameter settings. It is important to note that there is a trade-off between the window length and the ability to capture changes. Small windows will capture rapid changes but lose information (memory) about network stability. Within the scope of our work, this kind of window configuration is reflected in a set of network snapshots representing the state of the network for a sequence of fixed, typically short, time frames. Due to its forgetting mechanism, this approach provides a more up-to-date representation of the network, thus allowing to capture the most current events, which would otherwise be smoothed out by the whole historical data accommodated in a landmark window.

3.3 Literature Review on Dynamic Network Analysis

In real-world systems, the topology of most networks (*e.g.*, the WWW, co-authorship, biological, customer, and friendship networks) is temporal and dynamic in nature as networks tend to evolve gradually, through the addition and removal of nodes and links. This evolution is a result of changes

in the relationships or interactions among the network entities and manifests itself across the time axis. For instance, a new author may join the dynamic network in the co-authorship network, a new web page appears or disappears in the WWW, an individual starts a new social relationship or ends an existing one, a customer stops buying a given product, a blogger stops blogging, and a mobile network subscriber starts calling new people or reduces the communication with former contacts. Examining how the structure of these networks changes over time is an interesting problem as it potentially offers insights into both their evolutionary patterns and possible factors that trigger the changes, shedding light on the dynamic behaviour of the underlying system. The extracted knowledge can then be harnessed for predicting the future structure of these networks. In fact, a few studies have empirically demonstrated that explicitly incorporating the temporal behaviour of the network in the link prediction problem significantly improves the prediction results (Huang and Lin, 2009; Tylenda et al., 2009; Sarkar et al., 2012).

However, the great bulk of research in network analysis has, until recently, neglected the dynamic property of networks. The traditional approach discarded the temporal information by modelling the dynamic network as a static one, either by focusing the analysis on a snapshot of the network associated with a single point in time or on a larger network derived from the aggregation of all the interactions observed during the available time span. By overlooking a key characteristic of networks, research based on these strategies potentially missed valuable evolutionary patterns, or information about processes that are happening inside the network, that help explain its current configuration. A better and widespread strategy to study dynamic networks is to convert the evolving network into a temporally ordered sequence of static snapshots, each representing the state of the network at a given point in time.

In the last decade, research in dynamic network analysis has mainly proceeded in two different directions. These reflect different perceptions on what is an evolving network and on how the temporal data should be exploited for extracting knowledge about the underlying dynamics. The first thread of research views networks as complex systems, strives to seek laws that govern the evolution of a wide variety of networks, and aims at deriving a model that matches the discovered evolution laws. In turn, the second thread of research leverages the community structure of the network to both model the dynamics of the underlying network and identify significant changes in its structure. This second thread further divides into several categories that distinguish themselves by the kind of assumptions and mechanisms used to achieve these goals. Besides these two threads, we can also identify another strand of research, dubbed Dynamic Network Analysis (DNA), originating from the social sciences. DNA has been establishing roots as a separate field of research and can be understood as an extension of traditional SNA. According to Carley (2003), who is one of the most prominent researchers in this field, DNA aims at developing statistical methods and tools to analyse the evolution of large-scale, multimode (different types of nodes), multiplex (different types of ties), and multilevel (nodes in a network can be members of other nodes) networks. DNA

investigates networks using statistical analysis tailored for this specific problem and multi-agent simulations of network dynamics. For instance, the first attempts to study the evolution of networks modelled longitudinal network data as continuous-time Markov chains (Holland and Leinhardt, 1977; Wasserman, 1980). These models were further elaborated by Snijders et al. (2010) and are better known as *stochastic actor-based models*. In this literature review we will focus on the first two threads of research since these are the ones that influenced our work on network dynamics.

3.3.1 Models of Evolution Laws

Contributions to the first thread of research perform a retrospective analysis of different types of evolving networks (*e.g.*, the WWW, co-authorship networks, citation networks, biological networks, social interaction networks) by studying time series of several characteristics of the network (*e.g.*, degree distribution, density, effective diameter, average path length, size of the connected components) and the relationships among them. The multivariate time series analysis is based on the complete history of the network, which is assumed to be available *a priori*. Then, researchers look for key emerging evolutionary patterns shared by a wide range of networks in these time series, and derive a single model that is able to explain and replicate these patterns. These enduring patterns are often viewed as laws that govern the formation, growth, and evolution of real-world evolving networks. Some of the most important laws discovered so far in dynamic networks are the shrinking diameters, densification power-law, bursty weight additions, constant-size smaller components, and preferential attachment. These laws and the corresponding models (*e.g.*, preferential attachment model, edge copying model, forest fire model, community-guided attachment model) are presented and discussed in Leskovec et al. (2005), Leskovec et al. (2007b), Leskovec et al. (2008a), Barabási et al. (2009), Chakrabarti et al. (2010) and McGlohon et al. (2011). The models generated in this research thread are useful not only to understand general patterns in the network evolution that are generally true across a wide range of real networks, but also to build realistic generators of dynamic networks. Such generators are of utmost importance for validating the results of new algorithms (*e.g.*, community detection algorithms) and to help predict how the network will evolve, assuming that the underlying evolutionary process is stationary. However, these methods are more focused on finding general evolutionary laws and, thus, are not suitable for the detection of local changes occurring at specific regions of the network (*e.g.*, in a specific community of interest). Another drawback is related to the inability of these models to be easily updated upon the arrival of new data (nodes and links). In these cases, the whole analysis would need to be performed again and a new model created.

3.3.2 Community-based Models for Evolution Analysis

The second thread of research delivers a different type of knowledge about the evolution of a network by focusing the analysis on the associated community structure. Since networks are dynamic, substructures inside the network, namely communities, also change over time. Research advances on the analysis of evolving networks have been supported on the analysis of these substructures because the community structure is able to summarise and capture the broad patterns that characterise the whole network. This summarisation is relevant as it helps uncover existing patterns and predict emergent network properties (Gandomi and Haider, 2015). Therefore, detecting and quantifying changes in the community structure has been used as a strategy to both model and signal significant evolutionary events in the underlying evolving network. The models proposed in this thread of research share this general objective: to characterise the evolution of the network by monitoring the community structure over time. In contrast with the methods of the first thread, the produced models are incremental since they can be updated and adapted as new data arrive without the need to recompute the whole model. Another shared property of the majority of these methods is that they interpret a *dynamic community* as a pattern that can evolve in both membership and content, and manifests itself as a temporally ordered sequence of community instances that exist across a set of network snapshots over the time span. Despite these similarities, the distinct assumptions and approaches followed by these methods warrant a further categorisation. We thus categorise them according to their assumption on the stability/instability of the dynamic community structure.

Methods based on Evolutionary Clustering or Dynamic Probabilistic Modelling

Methods of the first category perform time-dependent dynamic community analysis and assume that the membership of communities may change but not drastically. Thus, these methods are suitable for analysing dynamic networks with stable community structures (*e.g.*, networks of social relationships). Prominent advances in this category encompass methods emanating from the *evolutionary clustering framework* (Chakrabarti et al., 2006) and methods relying on *dynamic probabilistic modelling*. Methods based on evolutionary clustering aim at finding communities in a particular snapshot that are meaningful and, at the same time, are temporally consistent with the communities found at previous snapshots. In other words, the goal is to maintain the freshness of the learned model by adapting it to currently observed moderate changes. The problem of discovering and monitoring communities is formulated as an optimisation problem that incorporates two potentially conflicting criteria: the snapshot quality and the temporal smoothness. The former guarantees the quality of the community model learned at each time step. The latter ensures that the community models are temporally similar, by penalizing dramatic shifts from previously learned models. Examples of methods inspired by the evolutionary clustering framework include the ones

developed by Chi et al. (2009), Kim and Han (2009), Gupta et al. (2011), and Wang et al. (2012). Variants of this framework include methods that do not solve the previous optimisation problem but still ensure temporal smoothness in the evolving community structure by applying incremental community detection algorithms (see, for instance, the work of Takaffoli et al. (2013)). Methods relying on dynamic probabilistic modelling also adhere to the temporal smoothness principle but differ on the definition of community. In this category, communities are viewed as latent concepts and, hence, each node may belong to different communities with a given probability. Examples of such methods are the ones proposed by Sarkar and Moore (2005), Lin et al. (2009), Fu et al. (2009), and Yang et al. (2011).

Event-based Methods for Monitoring Community Evolution

Methods of the second category address the problem of dynamic community analysis differently, by decomposing the problem in two independent stages: community discovery and community monitoring. Since the community structure associated with each snapshot of the evolving network is discovered independently, *i.e.*, without explicitly considering the temporal information and previous community structures, these methods reveal to be more appropriate for analysing networks with unstable community structures (*e.g.*, the WWW, "who-calls-whom" networks, online social networks). The emphasis of the majority of methods falling in this category is placed on the monitoring stage, where the goal is to trace the same community at distinct time steps and detect critical events in its evolution. In dynamic settings, communities may undergo a series of evolutionary events, such as growth, split, and disappearance, which characterise their *life-cycle*. For instance, a community discovered at time step t_i may separate into several communities in time step $t_{(i+1)}$, if the former community splits into two or more communities. The observation of this type of events in the community structure signals the occurrence of significant changes in the underlying network, thus providing an effective way to study the network dynamics (Aggarwal and Subbian, 2014). Since the majority of methods in this category rely on characterising the evolution of a given community in terms of a series of critical evolutionary events, hereinafter we refer to them as *event-based methods* or *event-based frameworks*. The main idea behind these methods was inspired by the seminal paper of Samtaney et al. (1994). This work reports a method for identifying and tracking coherent regions in 2D and 3D scalar and vector fields. To study the evolution of these regions over time, they presented certain evolutionary events for objects, such as continuation, creation, dissipation, bifurcation, and amalgamation. The information about evolution is pictorially represented using a directed acyclic graph. Although the research problem was different, a parallelism can be made between the three steps comprised in this approach and the stages of community discovery, community tracking, and community evolution visualisation of the more recent methods.

As previously mentioned, the event-based frameworks for community evolution analysis

encompass two main stages: community detection and community monitoring across the time span. The general procedure adopted by the majority of frameworks is as follows. In the first stage, communities are independently discovered in each snapshot of the evolving network using a static community detection algorithm. Each one of these snapshots depicts the relationships, or interactions, among the entities observed at a given time step. An exception is the framework proposed by Palla et al. (2007), because the tasks of community detection and community matching are performed jointly by applying the CPM community detection algorithm in an union graph. In the second stage, the discovered communities are tracked and matched across consecutive time steps based on their similarity. Given that communities are partitions, a straightforward and intuitive way to perform the matching is to use set theory and threshold-based rules to decide if two partitions are similar or not based on the size of their intersection. This similarity is often measured in terms of community membership, *i.e.*, by the proportion of nodes shared by two communities discovered at consecutive time steps. If two communities are detected as matches, then they are interpreted as being instances of the same community which spans over time. It is common to represent the results of the matching problem as a temporally ordered weighted multipartite graph, where nodes of each set represent communities detected at a given time step and there is an edge between pairs of communities found at consecutive time steps if they are considered as matches. These edges are often weighted by the similarity between the corresponding communities. Paths in the multipartite graph trace the evolution of a dynamic community. The stage of community monitoring also involves the identification of the critical evolutionary events experienced by the dynamic communities, which support the characterisation of their life-cycle. Additionally, some frameworks also introduce several metrics to measure the degree of community evolution and find interesting patterns, such as community stability, growth rate, and community lifetime.

In Tables 3.1 and 3.2 we list and summarise the main characteristics of the state-of-the-art event-based frameworks for community evolution analysis, in order to highlight their differences and similarities. Although originally tailored for monitoring changes in cluster structures, we also include the MONIC framework proposed by Spiliopoulou et al. (2006) because it can be easily adapted to the community setting. One of the distinguishing features of the MONIC framework is the incorporation of a data ageing function to accommodate the decay of older records and their importance in the cluster matching step. From the analysis of these tables, we can conclude that there is consensus among scholars regarding the type of evolutionary events a community may undergo during its life-cycle, though the terminology is not consistent. The tables also indicate that most of the methods assume crisp communities, use threshold-based rules for identifying communities' matches and rely on set theory measures (*e.g*, the Jaccard coefficient) for quantifying the similarity between pairs of community instances. Although we report the community detection algorithms used in the experiments carried out by the authors in the listed publications, most of these methods operate independently of the chosen static community detection algorithm. Another

Table 3.1: Summary of the main characteristics of the state-of-the-art event-based frameworks for community evolution analysis based on a two-stage approach. The publications are chronologically ordered. In the categories *Type of Window Model*, *Community Detection Algorithm* and *Type of Community* we consider the information reported by the authors in the experimental section of the corresponding publication. The acronym NA stands for Non-Available information.

Research Method	Community Detection Algorithm	Type of Community	Type of Window Model	Critical Evolutionary Events
Toyoda and Kitsuregawa (2003)	Companion– (a HITS-based related page algorithm)	Crisp communities	Landmark window	Growth, Shrinkage, Emergence, Dissolution, Split, Merge
Hopcroft et al. (2004)	Centroid-based agglomerative hierarchical clustering based on cosine similarity	Crisp communities	Landmark window	Established, Emerging
Mei and Zhai (2005)	Probabilistic mixture model estimated using the Expectation Maximization algorithm	Soft communities	Overlapping sliding window Non-overlapping sliding window	NA
Spiliopoulou et al. (2006)	Not Applicable	Crisp communities	Sliding window	Expansion, Shrinkage, Compaction, Diffusion, Survival, Emergence, Disappearance, Split, Absorption
Falkowski et al. (2006)	Girvan-Newman hierarchical divisive algorithm	Crisp communities	Overlapping sliding window	Growth, Decline, Split, Merge
Berger-Wolf and Saia (2006)	NA	Crisp communities	NA	NA
Palla et al. (2007)	CPM	Overlapping communities	Landmark window	Growth, Contraction, Birth, Death, Split, Merge
Asur et al. (2009)	NA	NA	Non-overlapping sliding window	Formation, Dissolution, Continuation, Split, Merge
Greene et al. (2010)	Louvain method	Crisp communities	Non-overlapping sliding window	Expansion, Contraction, Birth, Death, Split, Merge
Takaffoli et al. (2011)	Local community mining algorithm	Crisp communities	Non-overlapping sliding window	Formation, Dissolution, Survival, Split, Merge
Bródka et al. (2013)	Louvain method, CPM	Crisp and overlapping communities	Overlapping sliding window	Growth, Shrinkage, Formation, Dissolution, Split, Merge

Table 3.2: Summary of the main characteristics of the state-of-the-art event-based frameworks for community evolution analysis based on a two-stage approach (continuation of Table 3.1). The acronym NA stands for Non-Available information.

Research Method	Similarity Measure for Community Matching	User-defined Parameters	Community Evolution Measures	Community Evolution Visualisation
Toyoda and Kitsuregawa (2003)	$\max(\lvert C_t \cap C_{t+1} \rvert)$	None	Growth rate, Disappearance rate, Merge rate, Split rate, Stability, Novelty	Web community chart
Hopcroft et al. (2004)	$\min(\frac{\lvert C_t \cap C_{t+1} \rvert}{\lvert C_t \rvert}, \frac{\lvert C_t \cap C_{t+1} \rvert}{\lvert C_{t+1} \rvert})$	Fraction of trees out of n trees (f) Lower-bound for the best match (p)	NA	None
Mei and Zhai (2005)	Kullback-Leibler divergence	Evolution distance threshold (ξ)	NA	Evolution theme graph
Spiliopoulou et al. (2006)	$\frac{\sum_{a \in (C_t \cap C_{t+1})} age(a,t+1)}{\sum_{b \in C_t} age(b,t+1)}$	Matching threshold (τ_{match}) Split threshold (τ_{split})	Lifetime, Survival ratio, Absorption ratio	None
Falkowski et al. (2006)	$\frac{\lvert C_t \cap C_{t+1} \rvert}{\min(\lvert C_t \rvert, \lvert C_{t+1} \rvert)}$	Overlap threshold ($\tau_{overlap}$) Number of periods that may separate two community instances ($\tau_{periods}$)	Stability, Density and Cohesion, Euclidean distance, Correlation coefficient, Group activity	Graph of similar community instances
Berger-Wolf and Saia (2006)	$\frac{2\lvert C_t \cap C_{t+1} \rvert}{\lvert C_t \rvert + \lvert C_{t+1} \rvert}$	Turnover threshold (β) Persistence (α) Membership (γ)	NA	β-graph
Palla et al. (2007)	$\frac{\lvert C_t \cap C_{t+1} \rvert}{\lvert C_t \cup C_{t+1} \rvert}$	None	Stationarity	None
Asur et al. (2009)	Bit matrix operations	Merge/Split parameter (κ)	Popularity index	None
Greene et al. (2010)	$\frac{\lvert C_t \cap C_{t+1} \rvert}{\lvert C_t \cup C_{t+1} \rvert}$	Matching threshold (θ)	None	Evolution graph
Takaffoli et al. (2011)	$\frac{\lvert C_t \cap C_{t+1} \rvert}{\max(\lvert C_t \rvert, \lvert C_{t+1} \rvert)}$	Matching threshold (k)	Community lifetime, Members fluctuation	None
Bródka et al. (2013)	$\frac{\lvert C_t \cap C_{t+1} \rvert}{\lvert C_t \rvert} \cdot \frac{\sum_{x \in (C_t \cap C_{t+1})} SP_{C_t}(x)}{\sum_{x \in (C_t)} SP_{C_t}(x)}$	Formation/Dissolution parameter (fd), α, β	None	None

interesting conclusion derived from these tables is the fact that most authors report experiments using snapshots extracted under a sliding window scheme (overlapping or non-overlapping). Sliding windows are embedded with forgetting mechanisms that drop older data from the analysis thus allowing to uncover the most recent changes occurring in the network. The choice of the type of window model is often guided by domain knowledge of the dynamic network under analysis. Nevertheless, very few authors discuss the impact of this choice on the results of the community evolution analysis, with the exception of Mei and Zhai (2005), Greene et al. (2010), and a previous work of Bródka et al. (2013) published in 2012 (Saganowski et al., 2012). Mei and Zhai (2005) acknowledge that the choice of the length of the sliding window is not trivial. They argue that small window lengths allow to detect local temporal patterns due to the high resolution effect but hamper the ability to perceive overall evolution patterns. In turn, larger window lengths offer insights into the overall evolution trend but the introduced smoothing conceals the local dynamics. The view of Saganowski et al. (2012) is analogous and supported on empirical evidence obtained from extensive experiments. Greene et al. (2010) also underline the importance of the window length on the quality and stability of dynamic communities and encourage further research on this topic. In this chapter, we seize this research opportunity and address the impact of the window model choice in the context of a marketing application supported by real-world customer data.

Other methods for community evolution analysis, that are related to the ones discussed here, were proposed in the literature. Examples include parameter-free methods that use the minimum description length principle to both discover meaningful communities and meaningful change points in their evolution (Sun et al., 2007; Ferlez et al., 2008). For a comprehensive description of the methods discussed in this section and many others, we point the interested reader to the excellent surveys of Spiliopoulou (2011) and Aggarwal and Subbian (2014).

3.4 Methodology

In this section, we detail our methodology for analysing the structure of evolving customer networks, in terms of community evolution. The motivation for developing this methodology was driven by the need of companies to tap the potential of the customer-related data they have been accumulating over time. Here, we propose the use of social network analysis techniques and community evolution analysis methods to explore these data. The idea is to carry out an empirical analysis of the implicit customer network, in order to identify profiles of customers (given in the form of communities) and their evolution over a given time period. This information can be further explored for the design of marketing promotions tailored to each profile, for predicting future customer needs, and for the development of product recommender systems.

The proposed methodology is comprised of three main sequential steps that are independently applied to two different window models. The main steps are:

1. Analysis of the dynamic network, at both the network-level and node-level, using well known SNA measures, in order to obtain a concise description of the network topology;

2. Application of our extended event-based framework for dynamic community analysis, dubbed MECnet, to each window setting;

3. Interpretation of the dynamics of customers' communities (or profiles).

Finally, the results of these three main steps for each window model are compared.

3.4.1 Network Analysis

In the context of this work, an implicit customer network is modelled as an undirected weighted graph $G = (V, E, W)$, in which V denotes the set of n nodes, E denotes the set of m links, and W represents the weights of the edges in E (W is a non-negative matrix of size $n \times n$). Each node $i \in V(G)$ corresponds to a customer and each link e_{ij} denotes the co-purchasing behaviour of customers i and j. Each link has a non-negative weight w_{ij} associated with it, which indicates the number of co-purchased products by customers i and j. The higher this weight, the more similar the purchasing behaviour and preferences of customers i and j. Note that this customer network model emerges from the one-mode projection of the bipartite graph between customers and products. Thus, using the exact same data, the product network could also be analysed. Network analysis is performed on this network by first computing well known SNA measures (*e.g.*, degree centrality, betweenness centrality, density) and then interpreting their meaning within the context of customer networks. The interpretation of these measures provide us insight about the shape and structure of the network, as well as information regarding the role and centrality of each node in the network. These measures are usually divided according to the level of analysis one wants to perform: at the level of the basic entities (nodes) or at the level of the whole network. The former measures how a single node is embedded in a network from that single node's perspective. The latter computes how the overall network links are organised from the perspective of an observer that has a bird's eye view of the network. In this methodology we resort to both levels of analysis to get a description of the network topology. At the node-level, we analyse the following measures for undirected networks: degree centrality and betweenness centrality. At the network-level, we focus on density and modularity. Note however that the scope of application of our methodology is not restricted to these measures since others can be used.

3.4.2 MECnet for Tracking Community Evolution

MECnet is the term we use to refer to the extension of MEC to the community evolution setting. MEC can be easily extended to deal with communities, which is the equivalent of *clusters* for networks, due to its general nature and relative independence from the algorithm used to extract

clusters/communities. We say *relative independence*, in the sense that MEC is not restricted to a particular clustering algorithm, although it requires that the adopted algorithm partitions the data into disjoint groups by finding crisp clusters/communities (each object/node is assigned to exactly one cluster/community). MEC is a framework proposed by Oliveira and Gama (2012) to monitor the evolution of clusters. MEC traces evolution through the detection and categorisation of clusters transitions, such as births, splits, and merges. The clusters are extensionally defined, *i.e.*, each cluster is defined by the entities that were assigned to it by a given clustering algorithm. It takes as input a set of clusterings, each one generated at a different time step. It performs pairwise mappings, between clusters obtained at time step t_i ($i = 1, ..., T$, with T denoting the last analysed time step) and at a later consecutive time step $t_{i+\Delta t}$. The *mapping* process is based on the concept of conditional probability and is restricted by a user-defined threshold - the *matching threshold* τ, where $\tau \in [0.5, 1]$. Intuitively, this threshold indicates the proportion of mutual entities two cluster instances have to share in order for them to be considered instances of the same cluster. If $\tau < 1$, it is assumed that a cluster can be a match of other cluster even without keeping all of its members, thus allowing new members to join and existing members to leave.

Similarly to previously proposed frameworks (listed in Tables 3.1 and 3.2), MECnet is an event-based framework for identifying and monitoring dynamic communities that relies on a two-stage approach: (i) derive a model of the community structure at each time step and then (ii) adapt this model to the data arriving at the following time step and detect changes. More specifically, the first stage consists in independently discovering communities at each snapshot of the network using a static community detection algorithm. These snapshots represent the state of the network in a given time step, and these time steps are obtained through discretisation of the time axis. This kind of snapshot analysis can be very effective when dealing with *slowly* evolving networks, *i.e.*, networks in which the nodes or links are added or removed from the network at low rates (Aggarwal and Subbian, 2014). MECnet is not restricted to a specific static community detection algorithm, as long as the chosen algorithm partitions the network into disjoint communities, *i.e.*, identifies crisp communities (each node belongs to exactly one community at each time step). Our framework can also handle the addition of new nodes to the network, *i.e.*, the number of nodes does not have to be fixed over time. In the second stage, for each pair of consecutive snapshots, MECnet assesses the similarity among communities instances extracted at distinct time steps based on the proportion of mutual nodes shared, in order to identify correspondences among communities. This proportion evaluates the overlap of the communities' membership and is given in the form of a conditional probability, which is computed as follows:

$$\begin{aligned} sim(Com^m_{t_i}, Com^u_{t_{i+\Delta t}}) &= P(X \in Com^u_{t_{i+\Delta t}} | X \in Com^m_{t_i}) = \\ &= \frac{P(X \in Com^m_{t_i} \cap Com^u_{t_{i+\Delta t}})}{P(X \in Com^m_{t_i})}, \end{aligned} \tag{3.1}$$

where X is the set of entities (nodes) assigned to community $Com_{t_i}^{m}$ ($m = 1, ..., k_{t_i}$, with k_{t_i} being the number of communities returned by a given static community detection algorithm at time step t_i) and $P(X \in Com_{t_{i+\Delta t}}^{u} | X \in Com_{t_i}^{m})$ represents the probability of X belonging to community Com^u from $t_{i+\Delta t}$ ($u = 1, ..., k_{t_{i+\Delta t}}$, with $k_{t_{i+\Delta t}}$ denoting the number of communities extracted at $t_{i+\Delta t}$) knowing that X belongs to community Com^m obtained at a previous time step t_i.

Based on this information, MECnet then models the communities as nodes and their transitions as weighted edges, in an *evolution graph* (*i.e.*, a weighted multipartite directed graph). The weights of the edges correspond to the conditional probabilities $sim(Com_{t_i}^{m}, Com_{t_{i+\Delta t}}^{u})$, which are computed according to Equation (3.1). These weights indicate the probability of a transition from the source community at t_i to the target community at $t_{i+\Delta t}$. Each node set in the evolution graph represents the community structure of the network extracted at a given time t_i ($i = 1, ..., T$) and is denoted by ξ_i. The evolution graph represents the changes taking place within the community structure of the evolving network. The underlying conditional probabilities matrices of the evolution graph, along with the matching threshold τ, provide the essential information to perform many-to-many mappings between communities across consecutive time steps and allows the identification of the corresponding transitions. However, before detecting transitions, it is also necessary to conceptualise and formalise the changes likely to occur between communities. Several taxonomies for categorising the *transitions*, or *critical evolutionary events*, that a cluster/community, may experience during its life-cycle were proposed in the literature (Toyoda and Kitsuregawa, 2003; Spiliopoulou et al., 2006; Falkowski et al., 2006; Palla et al., 2007; Asur et al., 2009; Greene et al., 2010; Takaffoli et al., 2011; Gupta et al., 2011; Bródka et al., 2013). There is a general agreement among researchers on the type of critical evolutionary events that can be used to characterise the life-cycle of dynamic communities. There are only slight differences in the terminology used (*e.g.*, formation instead of birth, dissolution instead of death). Here, for simplicity, we adopt the taxonomy proposed in MEC (Oliveira and Gama, 2012). Thus, we consider that communities can undergo five different types of critical evolutionary events, or changes: *birth*, *merge*, *split*, *survival*, and *death*. These events can be briefly defined as follows:

- **Birth:** a new dynamic community appears as a result of the arrival of new entities into the network, or emerges by aggregating entities from older communities.
- **Death:** a previously discovered dynamic community disappears due to the loss of most of its constituent entities. Two possible reasons for the death of a dynamic community are: (i) the entities belonging to the community cease to exist or leave the network or (ii) the entities that made up the community become scattered across several dynamic communities at a later time step $t_{i+\Delta t}$.
- **Split:** a dynamic community from t_i splits into two or more distinct dynamic communities

in $t_{i+\Delta t}$ when a portion of its constituent entities disconnects, or separates, from the other constituent entities, giving rise to two or more smaller dynamic communities in $t_{i+\Delta t}$.

- **Merge:** two or more distinct dynamic communities from t_i fuse, or merge, into a single dynamic community in $t_{i+\Delta t}$ when the entities that belonged to the communities in t_i come together and create a larger community in $t_{i+\Delta t}$.

- **Survival:** a dynamic community survives when it does not undergo any of the above events. This is manifested by the presence of two identical community instances in consecutive time steps, which share at least 50% of entities (the lower bound of τ is 0.5, which implicitly assumes that a dynamic community persists only if its membership changes gradually).

Table 3.3: MECnet notation of critical community events.

Mathematical Notation	Description
$\emptyset \rightarrow Com^{u}_{t_{i+\Delta t}}$	Community Birth
$Com^{m}_{t_i} \rightarrow \emptyset$	Community Death
$Com^{m}_{t_i} \overset{\subseteq}{\rightarrow} \left\{Com^{1}_{t_{i+\Delta t}}, ..., Com^{r}_{t_{i+\Delta t}}\right\}$	Split of a community into r communities
$\left\{Com^{1}_{t_i}, ..., Com^{p}_{t_i}\right\} \overset{\subseteq}{\rightarrow} Com^{u}_{t_{i+\Delta t}}$	Merge of p communities into one community
$Com^{m}_{t_i} \rightarrow Com^{u}_{t_{i+\Delta t}}$	Community Survival

Table 3.4: Formulation of the critical evolutionary events a community may experience within the context of MECnet. Recall that τ denotes the matching threshold, whereas λ denotes the split threshold.

Events Taxonomy	Notation	Formal Definition
Community Birth	$\emptyset \rightarrow Com^{u}_{t_{i+\Delta t}}$	$0 < sim(Com^{m}_{t_i}, Com^{u}_{t_{i+\Delta t}}) < \tau \; \forall_m$
Community Death	$Com^{m}_{t_i} \rightarrow \emptyset$	$sim(Com^{m}_{t_i}, Com^{u}_{t_{i+\Delta t}}) < \lambda \; \forall_u$
Community Split	$Com^{m}_{t_i} \overset{\subseteq}{\rightarrow} \left\{Com^{1}_{t_{i+\Delta t}}, ..., Com^{r}_{t_{i+\Delta t}}\right\}$	$(\exists_u \exists_v : sim(Com^{m}_{t_i}, Com^{u}_{t_{i+\Delta t}}) \geq \lambda \wedge$ $sim(Com^{m}_{t_i}, Com^{v}_{t_{i+\Delta t}}) \geq \lambda) \wedge$ $\sum_{u=1}^{r} sim(Com^{m}_{t_i}, Com^{u}_{t_{i+\Delta t}}) \geq \tau$
Community Merge	$\left\{Com^{1}_{t_i}, ..., Com^{p}_{t_i}\right\} \overset{\subseteq}{\rightarrow} Com^{u}_{t_{i+\Delta t}}$	$(sim(Com^{m}_{t_i}, Com^{u}_{t_{i+\Delta t}}) \geq \tau) \wedge$ $\exists \, Com^{p} \in \xi_i \setminus \{Com^{m}\} : sim(Com^{p}_{t_i}, Com^{u}_{t_{i+\Delta t}}) \geq \tau$
Community Survival	$Com^{m}_{t_i} \rightarrow Com^{u}_{t_{i+\Delta t}}$	$(sim(Com^{m}_{t_i}, Com^{u}_{t_{i+\Delta t}}) \geq \tau) \wedge$ $\nexists \, Com^{p} \in \xi_i \setminus \{Com^{m}\} : sim(Com^{p}_{t_i}, Com^{u}_{t_{i+\Delta t}}) \geq \tau$

The adopted notation for critical community events is based on the one introduced by Spiliopoulou et al. (2006) and is given in Table 3.3. These events are external, as they relate to changes in the

whole community structure, and represent five basic types of changes that communities may undergo in consecutive time steps. A new threshold - the split threshold λ - was introduced in order to help distinguish community deaths from community splits. It is assumed that a community from t_i splits into, at least, two communities in $t_{i+\Delta t}$ if there are, at least, two communities $Com^u_{t_{i+\Delta t}}$ and $Com^v_{t_{i+\Delta t}}$, whose edges weights are equal or larger than the split threshold λ and the sum of these edges weights is equal or larger than the matching threshold τ. A formal definition of these events, given in the form of a set of rules, is provided in Table 3.4.

The proposed MECnet framework is illustrated in Figure 3.1 through a toy example. This example features three network snapshots, that represent the state of a dynamic network at consecutive moments in time, and the corresponding communities uncovered by a given static community detection algorithm. The chosen algorithm is independently applied to each snapshot of the network in order to discover the community structure associated with a given time step. The number of detected communities is 4 for all time steps under analysis: $k_{t_1} = 4$, $k_{t_2} = 4$, and $k_{t_3} = 4$. These community structures are represented as disjoint sets of 4 communities in the evolution graph. The dynamics of the communities are then studied using the MECnet framework. For instance, the red and blue communities discovered at time step t_1 merge into the purple community at time step t_2. Since the weights of the edges associated with the red and blue communities is equal to 1, this merge occurs by preserving all the entities of these communities. In the following time step t_3, the purple community survives, as can be ascertained from the evolution graph. The information provided by MECnet regarding the evolution of the red community can be summarised as follows: the temporal trajectory of the red dynamic community of time step t_1 can be described as $TT_{Com^{red}} = \left\{Com^{red}_{t_1}, Com^{purple}_{t_2}, Com^{purple}_{t_3}\right\}$, and its life-cycle until time step t_3 is given by $LC_{Com^{red}} = \left\{born_{[t_1]}, merge_{[t_1,t_2]}, survival_{[t_2,t_3]}\right\}$. A similar reasoning is used to describe the evolution of the remaining communities of this example.

3.4.3 Window Models

Our methodology makes use of two well known window models to analyse the dynamics of the network at the community-level: the landmark window and the overlapping sliding window. Although we have also carried out experiments with non-overlapping sliding windows of size 90 days, we considered that the obtained results were not interesting for our application due to the loss of information regarding the temporal continuity of customer communities and consequent difficulty in finding stable customer profiles. In other words, there were too few customers buying in consecutive time frames, which reflected in a very irregular behaviour of the network. The specific reasons for discarding the non-overlapping sliding window from our application-driven methodology are related to (i) the nature of the business of the company under analysis, which operates in the Business-to-Business (B2B) market, with the great majority of customers being other companies; and (ii) the nature of the products sold by the company, which have long life-cycles and

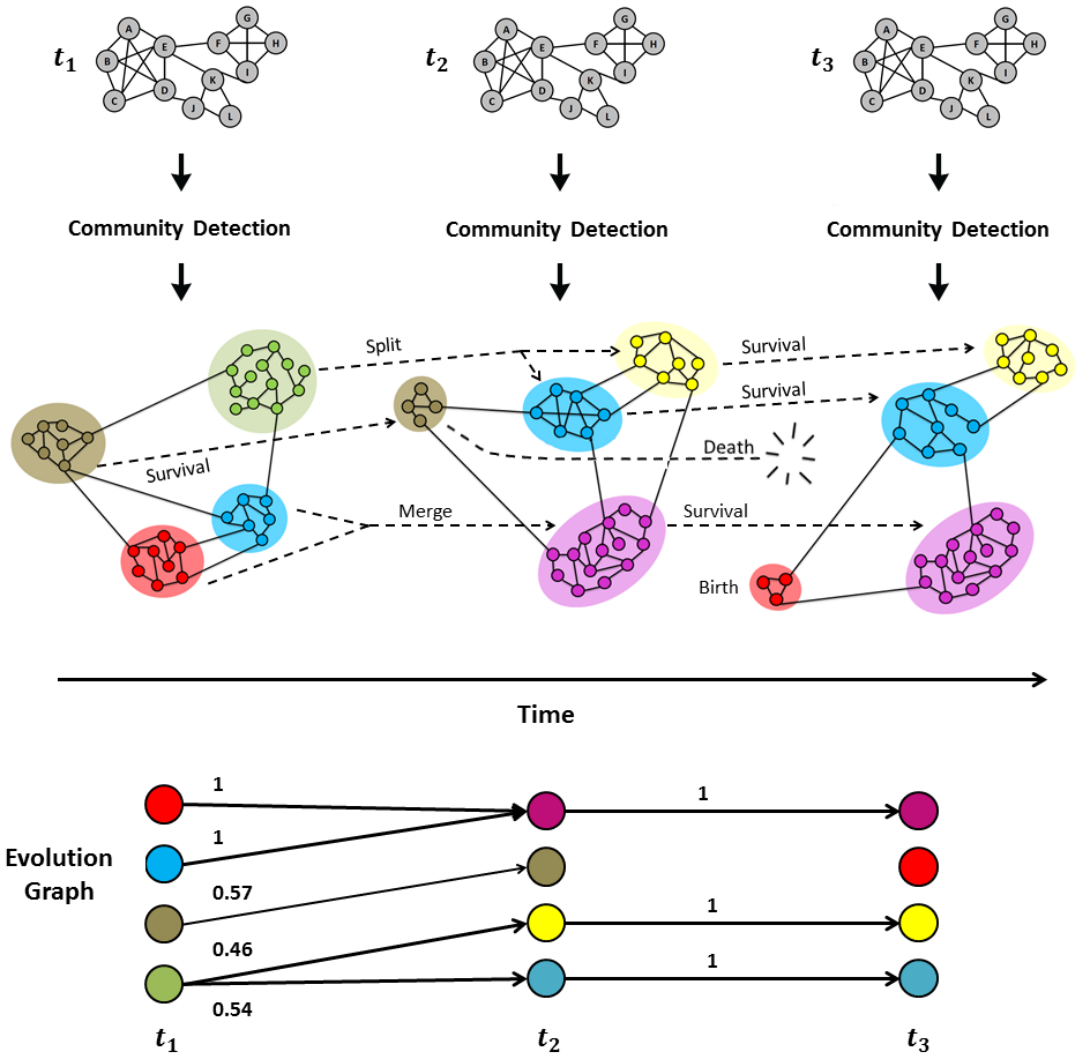
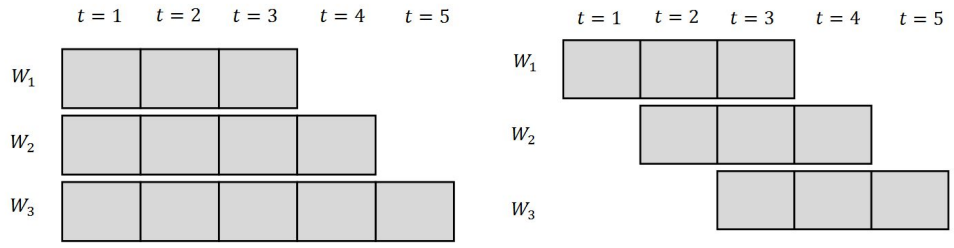


Figure 3.1: Illustration of the MECnet framework using an example of a dynamic network and its evolution through time span $[t_1, t_3]$. The evolution graph summarises these dynamics and captures several types of critical evolutionary events, namely: birth, split, merge, survival, and death.

a low purchase frequency (please see Section 3.5.1 for a more detailed description of the company's products). In this scenario, the non-overlapping sliding window model proved of limited usefulness, since it only confirmed what was already known, thus not being able to provide actionable insights into the evolution of customer profiles. Note, however, that these conclusions only hold for these specific customer data. Therefore, if this methodology were to be applied to other types of customer data, the non-overlapping sliding window model could eventually provide relevant information.

In this section, we explain how we formulate this problem for each type of window model considered in this study, on the context of MECnet.



(a) **Landmark window** (b) **Sliding window**

Figure 3.2: Illustration of (a) a landmark window, and (b) a sliding window with length three time steps and a step width of one time step.

Landmark Window

The dynamic customer network is modelled as a temporally ordered sequence of T graph snapshots $\{G_1, G_2, ..., G_T\}$, where $G_i = (V_i, E_i, W_i)$ represents a static cumulative snapshot of the network at a given discrete time step t_i $(i = 1, ..., T)$, depicting all the nodes and links observed up to the current time step (*e.g.*, G_3 comprises all the nodes and edges observed during the time span $[t_1, t_3]$). The landmark window approach successively aggregates these static snapshots into a unique graph, as illustrated in Figure 3.2 - (a). MECnet is then applied to the accumulated network observed at each time step by:

1. First detecting the communities using a static community detection algorithm (*e.g.*, Louvain method, Label Propagation algorithm) and then,

2. Modelling the evolution of these communities, for a sequence of T time windows, through an evolution graph.

The k_{t_i} communities found at time step t_i are denoted as $Com_{t_i}^m$ $(m = 1, ..., k_{t_i})$. MECnet allows the characterisation of the life-cycle of each dynamic community (for example, a dynamic community

is born in t_1, survives in t_2, merges with other two communities in t_3, survives in t_4, t_5, t_6, and t_7, splits in t_8, and dies in t_9). The communities found at each time step are referred to as *instances*, or *observations* of a dynamic community or, alternatively, as *step communities* (Falkowski et al., 2006; Greene et al., 2010). A dynamic community (also referred to as *meta-community* or *meta-group* in Berger-Wolf and Saia (2006); Takaffoli et al. (2011, 2013)) is described by a sequence of similar community instances, which define its *temporal trajectory* (also denoted as *timelines* by Greene et al. (2010)). The life-cycle of a dynamic community is defined as a temporal sequence of critical evolutionary events undergone by the dynamic community, from the moment it first appeared until the moment it fade away.

Sliding Window

The dynamic customer network is also modelled as a temporally ordered sequence of T graph snapshots $\{G_1, G_2, ..., G_T\}$, where $G_i = (V_i, E_i, W_i)$ represents a static snapshot of the network at a given discrete time step t_i $(i = 1, ..., T)$. In contrast with the landmark window, where all static graph snapshots are accumulated over time, in the sliding window approach only a pre-defined number of static graph snapshots are considered for the temporal analysis. We propose the use of an *overlapping* sliding window in order to guarantee that there is always a mutual time step between consecutive instances of the window. This condition prevents highly disruptive transitions, which are unlikely in real-world networks (Tantipathananandh et al., 2007). The overlapping sliding window approach first partitions the time axis into time slots of fixed length w and then it employs a forgetting mechanism by considering only the static graphs falling within each one of these slots. Thus, whenever a graph G_i is observed and inserted in the window, another graph G_{i-w} ($i > w$ and $w < T$) is forgotten. Such catastrophic forgetting allows us to focus only on current events, by considering in the analysis only the most recent nodes and links of the dynamic network. Community evolution is then studied by applying MECnet to the set of time windows (*e.g.*, time windows G_{1-3}, G_{2-4}, G_{3-5} for a window length $w = 3$ and a step width of one time step). In Figure 3.2-(b) we illustrate three time steps (W_k, $k = 1, ..., 3$) of an overlapping sliding window of length three time steps ($w = 3$) and step width of one time step.

3.5 Case Study

In this section, we proceed to validate the feasibility of our methodology using a real-world customer network, extracted from one of the largest Portuguese groups operating in the electric field. The goal of the company was to use, for the first time, SNA techniques to perform an empirical dynamic analysis of their customers' purchasing behaviour, so as to identify differentiable customer profiles (or customer communities) and their evolution over a given year. Since it is known that some customers are frequent buyers, whereas others engage in more sporadic purchases, we considered

relevant the analysis of the community dynamics using two distinct time window models, in an attempt to capture the behaviour of both types.

3.5.1 Network Data

The network data are imported from the company's CRM application and corresponds to a time span of twelve months (year 2011). We model the network based on the similarity of the purchasing behaviour between customers of the company. Thus, each customer is represented as a node in the network and there is a link between a pair of customers in t_i ($i = 1, ..., 12$) if they both purchased the same product during t_i. This link is weighted by the number of co-purchased products. The resulting network is undirected and weighted. For the chosen time span, the company's product portfolio was comprised of nearly 200 different products, from which 152 products were actually bought by the set of customers under analysis. The company's main products are related to the electric field and, thus, some degree of technological evolution can be observed. These products are typically supported by other products sold by the company in the form of service/maintenance contracts. The company also sells additional products related to engineering and high-tech projects. Two distinctive characteristics of these products are their long lifespan and low purchase frequency. The products' nature has impact on the dynamics of customer profiles and, consequently, on the number and type of detected events. The total number of nodes (active customers) and links in the whole network G_{1-12} is 1014 and 12259, respectively. The manipulation, visualisation, and analysis of the network is performed on Gephi (Bastian et al., 2009) by making use of its dynamic network analysis features. The community evolution analysis is mostly performed on `R` environment (Team, 2008), since the MECnet framework was implemented in this scripting language.

Due to confidentiality issues, in this case study we do not report detailed results regarding the business characterisation of the dynamic communities (type of customers, purchasing amount, characteristics of the products, etc.) and our interpretation is mostly performed at the macro-level.

3.5.2 Experimental Setting

We apply our methodology to the customer network by sequentially following the steps outlined in Section 3.4.

Instead of analysing a single aggregated snapshot of the entire available network G_{1-12} (static analysis at the macroscopic level), we explore the dynamics of the network at a mesoscopic level (*i.e.*, community-level), by making use of two window approaches. For the landmark model, we start with a window of three months and then we cumulatively grow the window by adding one month at each step. For the overlapping sliding model, we set the window length to three months ($w = 3$) and its step width to one month. In both cases, the total number of time steps is ten. The window length and step width were set by the company's Business Intelligence analyst, based on

his business knowledge regarding, *e.g.*, the rate of change in the customer network. Each time step, denoted as W_k ($k = 1, ..., 10$), is a time interval starting at t_i and ending at t_{i+w}. The time steps for the landmark window are: $W_1 = [t_1, t_3]$, $W_2 = [t_1, t_4]$, $W_3 = [t_1, t_5]$, $W_4 = [t_1, t_6]$, $W_5 = [t_1, t_7]$, $W_6 = [t_1, t_8]$, $W_7 = [t_1, t_9]$, $W_8 = [t_1, t_{10}]$, $W_9 = [t_1, t_{11}]$, and $W_{10} = [t_1, t_{12}]$. The time steps for the sliding window are: $W_1 = [t_1, t_3]$, $W_2 = [t_2, t_4]$, $W_3 = [t_3, t_5]$, $W_4 = [t_4, t_6]$, $W_5 = [t_5, t_7]$, $W_6 = [t_6, t_8]$, $W_7 = [t_7, t_9]$, $W_8 = [t_8, t_{10}]$, $W_9 = [t_9, t_{11}]$, and $W_{10} = [t_{10}, t_{12}]$. We detect the communities at each W_k using the Louvain method (Blondel et al., 2008) since it produces disjoint partitions of good quality, in a very fast way. Next, we apply MECnet to identify the critical evolutionary events undergone by the found communities. We set the *matching threshold* of MECnet to $\tau = 0.5$ and the events are detected at consecutive snapshots of the network (*i.e.*, time step intervals $[W_k, W_{k+1}]$). The reason for choosing a low value for τ is related to the low purchasing frequency of the company's products and the consequent need to ensure a reasonable persistence of customer communities. This choice of the matching threshold value was also supported by the results of the sensitivity analysis presented in Section 3.5.4. Finally, we evaluate both approaches based on a double perspective. First, we compute a quantitative measure, namely the *Survival Ratio* proposed by Spiliopoulou et al. (2006), to measure network volatility and the frequency of community transitions in both scenarios. This ratio is given in Equation (3.2) and it computes the portion of communities found at time step W_k ($k = 1, ..., 10$) that survived in W_{k+1}.

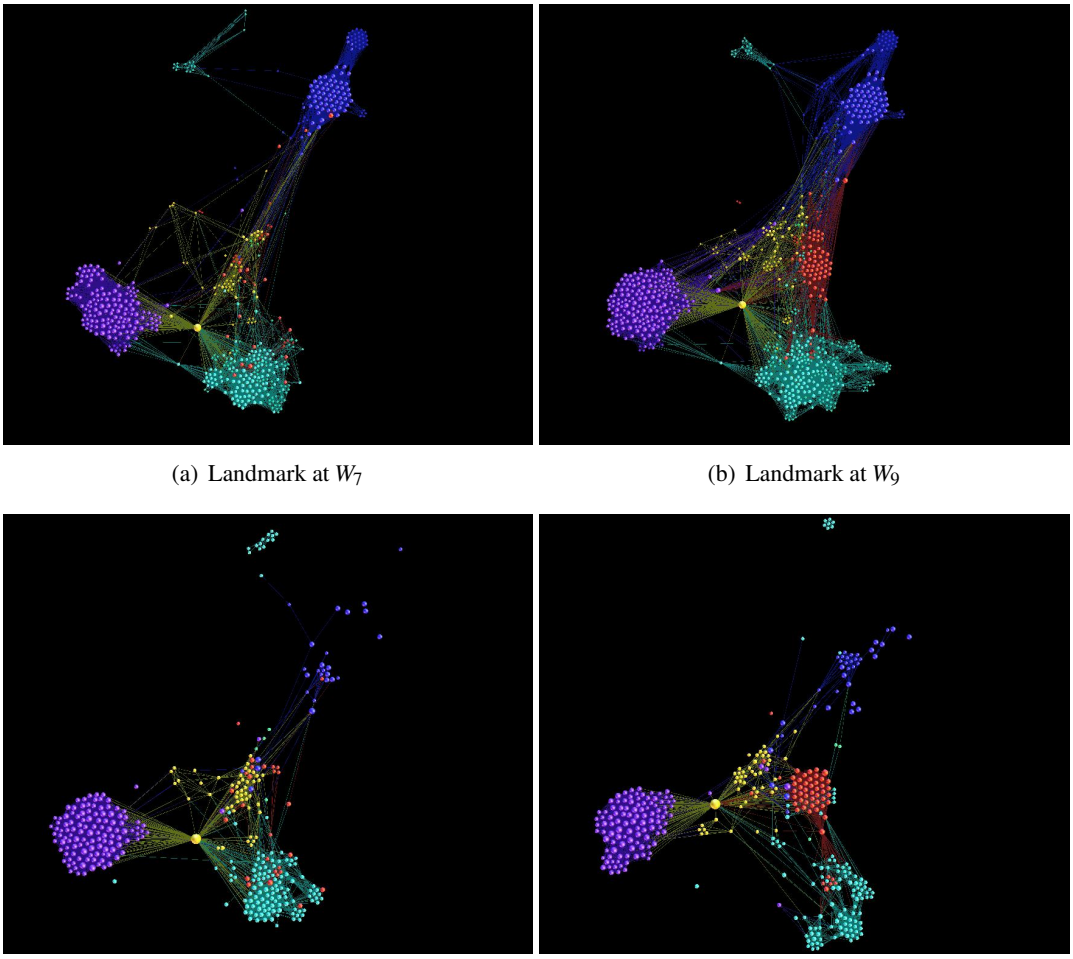
$$SurvivalRatio(W_k) = \frac{\#SurvivedCommunities(W_{k+1})}{\#Communities(W_k)}. \tag{3.2}$$

Then, we qualitatively assess the actionable insights derived from the analysis of each window model, from the business viewpoint.

3.5.3 Results

Following the experimental procedure described before, we obtain two dynamic networks. In Figure 3.3 we show two snapshots of each one of these dynamic networks, for time steps W_7 and W_9. For the first scenario (Figure 3.3-(a)/(b)), the number of nodes varies between 336 (first time step) and 1014 (last time step), whereas the number of links ranges from 3509 to 12259 (please see Figure 3.4). In contrast, in the second scenario (Figure 3.3-(c)/(d)), the number of nodes and links is more unstable over time, ranging from 227 to 336 customers and from 1519 to 3509 links, as can be ascertained from Figure 3.4. This is explained by the forgetting mechanism employed by the sliding window model.

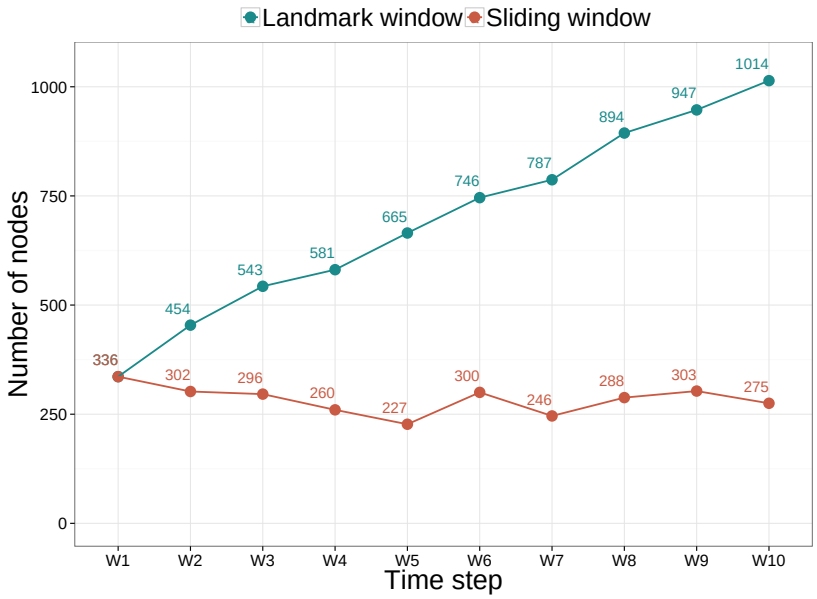
We compute the following SNA measures, which were considered to be of higher relevance within the scope of this case study: degree centrality, betweenness centrality, density, and modularity. Since the former two are node-level measures and, thus, need to be computed for each node, for simplicity we only present their average. The meaning of these measures within the scope of our



(a) Landmark at W_7 (b) Landmark at W_9

(c) Sliding at W_7 (d) Sliding at W_9

Figure 3.3: Two snapshots of a dynamic customer network. The top figures represent the network generated by the landmark window model at two distinct time steps (W_7 and W_9). The bottom graphs represent the network obtained using a sliding window model for the same time steps. The communities detected by the Louvain method are identified by different colours.



(a) **Number of Nodes**

(b) **Number of Links**

Figure 3.4: Number of nodes (n) and number of links (m) of each network snapshot, for the landmark and the sliding window models.

application is as follows:

- *Degree Centrality:* high values of the degree centrality are associated with customers who either buy the best-selling products of the company, or a wide variety of products. The former correspond to *single-product customers* and the latter to *multi-product customers*. The first type of customer is embedded in large network cliques. These network cliques are formed by sets of customers who buy the same product or products of the same type. If the clique is large, then the associated product is popular and generates high revenues for the company. The second type of customer are likely to be loyal and satisfied with the company, as they purchase different products from the company's product portfolio. Multi-product customers are valuable for the company and typically score high on both degree and betweenness centralities, whereas single-product customers score high on both degree centrality and clustering coefficient. Although the identification of single-product and multi-product customers is not within the scope of this chapter, companies can easily identify them in the network by computing joint rankings of node-level measures (degree centrality and betweenness centrality; degree centrality and clustering coefficient). The average degree centrality is not as informative as the degree centrality, but it is able to summarise in a single score the similarity of the co-purchasing behaviour of customers. Thus, high values of this measure reflect a high similarity in the customers' purchasing behaviour, whereas low values are associated with a more idiosyncratic purchasing behaviour among the company's customers.

- *Betweenness Centrality:* customers with high betweenness centrality occupy gatekeeper positions in the network. When averaged over all customers, betweenness centrality gives us an idea of the prevalence of such gatekeepers in the network. These gatekeepers tend to exhibit a distinctive purchasing behaviour, as they are the ones who usually buy the high value-added products of the company, or products belonging to very different categories. Intuitively, large scores of betweenness centrality are associated with customers who are loyal, satisfied or engaged with the company, since they buy a wide range of products from the company's product portfolio. The temporal analysis of the average betweenness centrality indicates the prevalence of such customers in the network and can help in the identification of trends in the customer network (*e.g.*, diversification, change of technology) which, in turn, can help unveil overall trends in the market itself.

- *Density:* measure of the network connectedness level. When taking a dynamic view of the network, a high density reveals a certain "maturation" of the network, both in terms of customers and their purchasing behaviour. On the other hand, a sparser network indicates a less mature network, since customers exhibit different product preferences.

- *Modularity Q:* quality function that attempts to measure the merit of a given partition of the network into communities. Measures the difference between the number of within-community edges and between-community edges in a given set of communities, and the expected number of within-community edges and between-community edges in a random network with the same order and the same average degree centrality. Large modularity values ($Q \geq 0.3$) suggest the presence of meaningful community structures (Clauset et al., 2004), which is the equivalent to meaningful customer purchasing profiles in our application.

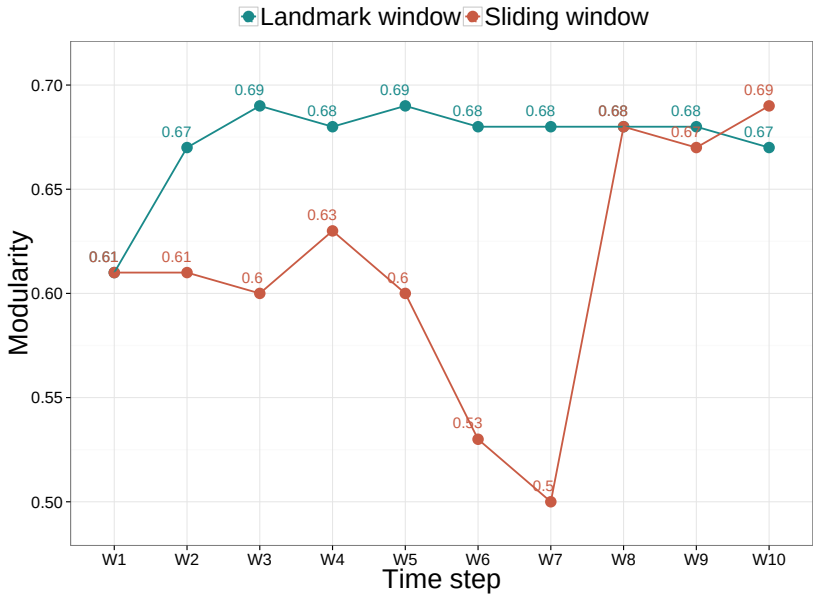
The obtained values are presented in Figures 3.5 and 3.6. Regarding the landmark window, the degree centrality averaged over all customers in the network exhibits a weak growing trend, starting at 20.89 (in W_1) and ending at 24.18 (in W_{10}). This result makes sense because, as the network grows, new customers join the network and are likely to buy products that were already bought by older customers, thus contributing to the increase of the average degree centrality. However, this increase is not significant, as it corresponds to an increment of approximately three links over a time span of about seven months. The average degree values are themselves low when compared to the number of customers, suggesting a relatively low level of co-purchasing similarity among the company's customers. This is reflected in a low network density and a relatively large number of communities. Concerning betweenness centrality, its average value increases as the window covers more time intervals, due to the existence of a few temporally consistent hubs linking distinct communities. Since the network is growing cumulatively, this consistent increase in the average betweenness centrality emphasises the strengthening of the gatekeeper positions of a few customers, which exhibit distinctive purchasing behaviours. In what regards density, the low values indicate a sparse dynamic network, with different sets of customers exhibiting distinctive buying preferences. We also observe a decrease in density over time, which might be explained by a non-linear increase in both customers and links. These values reveal that the customer network under analysis might not yet be mature and, thus, still has space for growth.

Concerning the sliding window, we observe an overall instability in the values of the measures. This was expected due to the time-based forgetting mechanism incorporated in the model. The fluctuation of the average degree centrality suggests a lack of customers' purchasing activity. This is related to the company's business nature, which relies on long-term projects and high value-added products, which in turn are typically associated with low-frequency purchases. Once again, the density of the network is close to zero due to the sparsity of links between customers.

Evolution of Customer Profiles

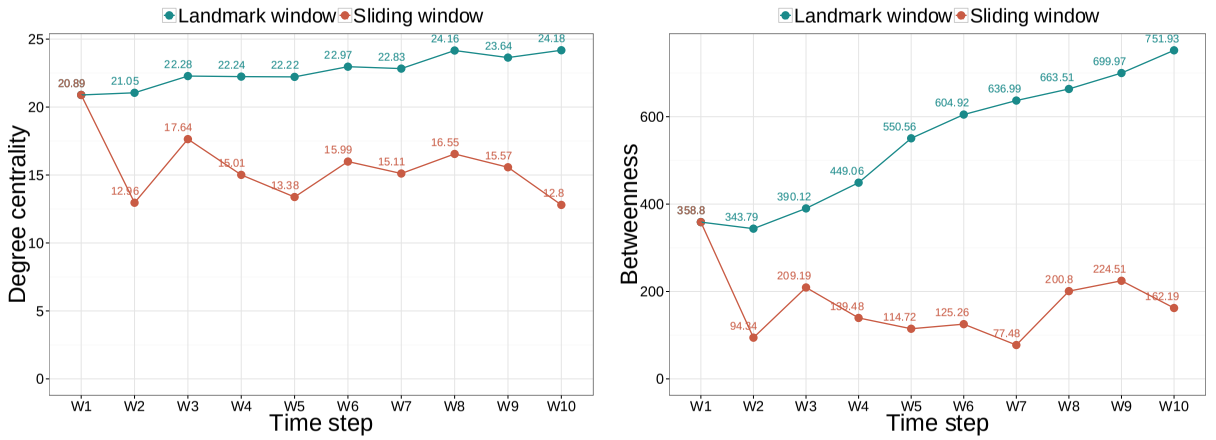
MECnet detects several types of events that mirror the dynamics of customer profiles (or communities). In this chapter, we focus only on survivals, births, and deaths, since these were found to be more representative of the dynamics occurring at the community-level. Given the relatively

(a) **Number of Communities**



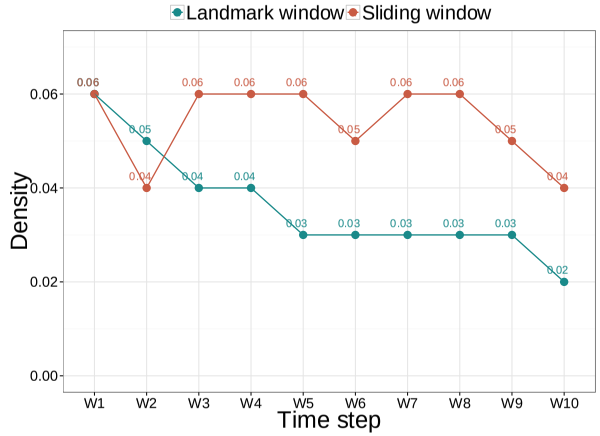
(b) **Modularity**

Figure 3.5: Number of communities returned by the Louvain method at each time step W_k ($k = 1, \ldots, 10$) of the landmark and the sliding window models, and the corresponding modularity.



(a) **Degree Centrality**

(b) **Betweenness Centrality**



(c) **Density**

Figure 3.6: Values of the node- and network-level measures, for each time step W_k ($k = 1, \ldots, 10$) of the landmark and the sliding window models. The node-level measures (degree and betweenness centralities) were averaged over all nodes.

large number of communities detected by the Louvain method at each time step (the initial number of communities ranged from 22 to 25, for different time steps and for both window models), we restricted our analysis to the number of communities representing, at least, 75% of the total number of customers in each network snapshot. The motivation for imposing a threshold on the representativeness of communities was to ensure that communities formed by isolated nodes, or by a small number of nodes, were discarded from the analysis. From a business viewpoint, this choice warrants the identification of a more manageable number of customer profiles while helping the company focus on the most representative purchasing patterns of its customers' universe. This way, the company is able to focus the analysis on those customers who buy their core products, while helping the marketing team wisely manage the resources (time and money) needed to improve the customer relationship. The minimum threshold was set by the Business Intelligence analyst and its choice was guided by his business knowledge. In our experiments, the representativeness of communities ranges from 76% to 85%. The effect of using this threshold is a drop in the number of communities. In our experiments, we obtain a minimum of 8 and a maximum of 12 communities, for the landmark window model, and 9 communities for the sliding window model (please see Figure 3.5-(a)). The average modularity Q was 0.61 for the sliding window, and 0.67 for the landmark window, which suggests the existence of well defined and meaningful communities of customers (please see Figure 3.5-(b)).

Due to its greedy nature, the Louvain method is not completely deterministic. If we change the order of the nodes, this method might return different results. This variability in the outcome is partly explained by the sequential nature of the analysis of modularity gains, which makes the method highly dependent on the starting node. Bearing this in mind, we evaluated the stability of the results of the Louvain method, in terms of modularity, number of communities, and Normalised Mutual Information (NMI) (Danon et al., 2005), by first shuffling the nodes IDs and then running the algorithm on the resulting network. The NMI, denoted as $I(A,B)$, is a well known measure of similarity borrowed from information theory, which has proved to be reliable in comparing network partitions (Lancichinetti and Fortunato, 2009). The closer NMI is to 1, the more similar the two network partitions are. In turn, if the two network partitions A and B are totally independent, $I(A,B)=0$. We compute the NMI between network partitions A and B as follows:

$$I(A,B) = \frac{-2\sum_{i=1}^{k_A}\sum_{j=1}^{k_B} N_{ij}\log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{k_A} N_{i.}\log(\frac{N_{i.}}{N}) + \sum_{j=1}^{k_B} N_{.j}\log(\frac{N_{.j}}{N})}, \tag{3.3}$$

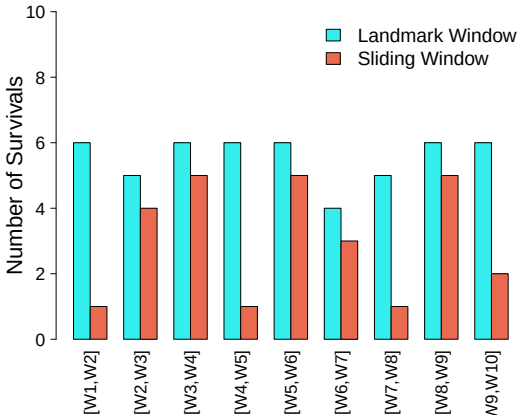
where $\mathbf{N}$ is a *confusion matrix*, in which the rows correspond to the k_A communities associated with network partition A and the columns correspond to the k_B communities associated with network partition B. The matrix elements N_{ij} denote the number of nodes in the community i of network partition A that also appear in community j of network partition B. The sum over row i of matrix $\mathbf{N}$ is denoted as $N_{i.}$ and the sum over column j is denoted $N_{.j}$.

We resort to the NMI to compare a baseline community structure, *i.e.*, the partition returned by the Louvain method without shuffling the nodes IDs (network partition A), with the community structure returned by each run of the Louvain method after randomly shuffling the nodes IDs (network partition B). We perform our evaluation using the network corresponding to time window G_{1-3}, which is exactly the same for both window approaches. For 50 runs of the algorithm (*i.e.*, 50 shuffles of nodes IDs), the average modularity was 0.67 ($\sigma = 0.002$), the average number of communities, without considering the 75% threshold, was 23 ($\sigma = 1$), and the average NMI was 0.93 ($\sigma = 0.03$). These values suggest a high stability in the outcomes of the Louvain method for the analysed network, since we obtain similar community structures, as revealed by the large NMI value and by the identical number of communities, without compromising their quality, as indicated by the large modularity value. The difference in the average modularity value obtained in this experiment and the one reported for time step W_1 in Figure 3.5-(b) is explained by the fact we are considering all the communities returned by the Louvain method and not only those that satisfy the 75% threshold.

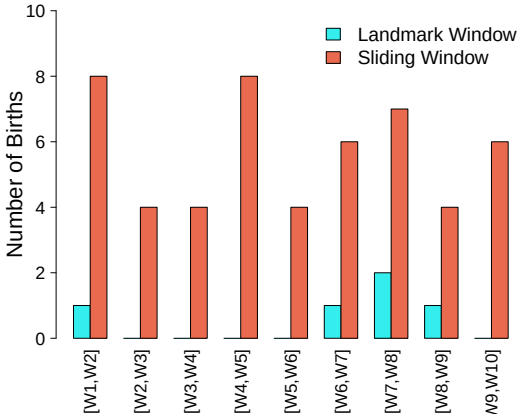
The total number of survival, birth, and death events detected by MECnet (with a matching threshold of $\tau = 0.5$) for both window approaches, at each time step interval, is provided in Figure 3.7. From the analysis of these figures, we can deduce that the overlapping sliding window model is able to capture the unstable community structure of the customer network, by focusing only on the most recent past. This volatility is reflected on the high number of births and deaths, on the relatively low number of survivals, and is captured by the values of the survival ratio (please see Figures 3.7 and 3.8). These differences on the number of events show the potential of the sliding window in capturing temporary acute changes occurring in the network, reflecting more closely the changes in customers' preferences for products. At the community-level, these changes manifest themselves through sequences of death and birth events. Please note that nodes that are forgotten in the sliding window approach might re-appear at later temporal snapshots. In these cases, they are assumed as newborn nodes. Regarding the landmark window, the large number of survivals, as captured by the survival ratio, suggests little volatility between time steps (Figure 3.8). The few birth/death detected events are related to more stable and long-lasting changes at the customer network.

Qualitative Evaluation and Discussion

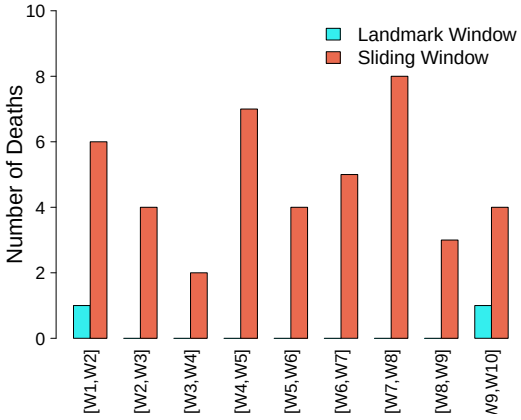
Cliques: our definition of customer network implies that customers who buy the same product in a given timeframe are all linked to each other. Thus, it is likely that cliques are formed between customers who buy the same products. A closer look at the customer networks provides empirical evidence that supports this hypothesis. In fact, most members of a given community bought the same product or, alternatively, bought products of the same type. This is corroborated by the analysis of the global clustering coefficient of the network, which is consistently higher than 0.8,



(a) **Survivals**



(b) **Births**



(c) **Deaths**

Figure 3.7: Events detected by MECnet (matching threshold $\tau = 0.5$) for the landmark and sliding window approaches.

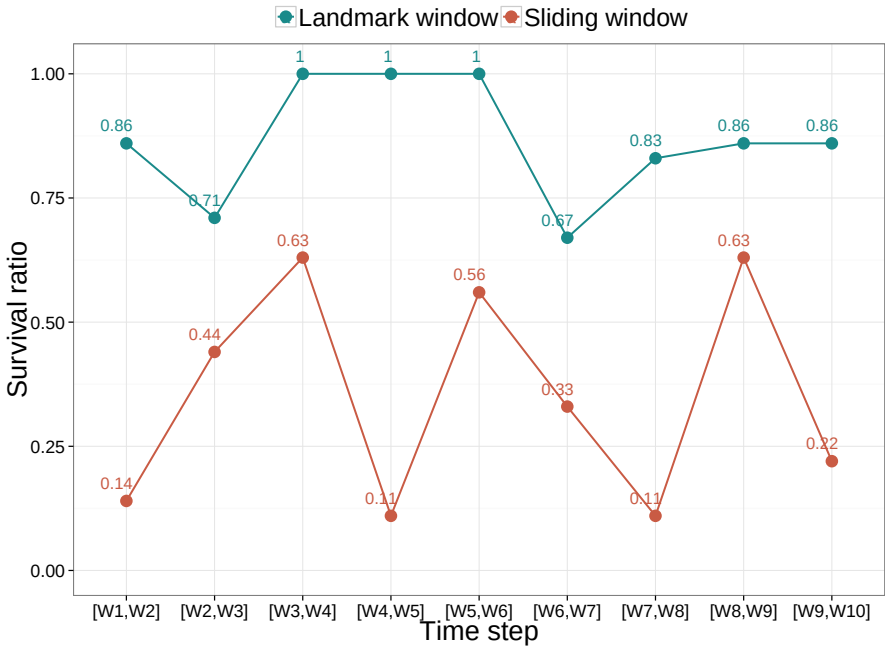


Figure 3.8: Survival ratio for the landmark and sliding window approaches.

regardless of the window model considered. This suggests that most communities are themselves cliques or formed around cliques. These findings suit our initial purpose of identifying customers with similar profiles in terms of products' preferences. On the other hand, when taking into account the temporal dimension of these networks, it can be hypothesised that these customers' cliques shift over time from the usage of a given product to the usage of another product. Due to the nature of this company's business, its products' characteristics, and the availability of data for a single year, it is not possible to properly test this hypothesis. However, this shifting behaviour can be observed at the micro-level for several customers, signalling the possibility of observing this behaviour for whole communities in the case we had access to a longer time frame of data.

Effect of the network model on the results: we model the implicit relationships among customers based on the similarity of their buying behaviour. In this kind of network model the type of products bought by these customers plays an important role on both the network structure and the nature of the detected events. The diversity of the company's product portfolio, in terms of products' categories, products' function, and price ranges, is able to meet the needs of different types of customers, thus reflecting in a more idiosyncratic purchasing behaviour. This distinctive behaviour, shared by different sets of customers, partly explains the existence of well defined customer profiles (*i.e.*, communities) revealed by the large modularity values. Furthermore, the long life-cycle of products (typically, more than 10 years) reflects in low replacement rates and low

purchase frequencies. According to the company's data, 50% of customers made a single purchase, whereas 17% made only two purchases during 2011. This explains the high number of births and deaths of communities when using the overlapping sliding window model. Exceptions to this rule are customers that are large companies. These large corporations usually engage in repeated purchase, since they buy a product several times, as well as its associated products (*e.g.*, service contracts). On the other hand, the merges, the splits, and the individual migration of customers from one community to another, can be partly explained by the associated products the company sells (*e.g.*, the purchase of a typical electric product is usually followed by the acquisition of a maintenance/service contract). Based on this qualitative analysis, it became clear that the choice of the network model influences the obtained results, so a careful study of the most appropriate model should precede the application of the proposed methodology.

Stability of the static community detection algorithms: an important issue in dynamic community analysis is related to the stability of the static community detection algorithms used for discovering community instances. Most state-of-the-art community detection methods are surprisingly sensitive to minor modifications of the input networks. Therefore, small perturbations of the input data may lead to significant changes in the resulting community partition. This can be problematic when finding correspondences between two independently generated partitions, since it becomes hard to distinguish between changes driven by the network evolution or by the instability of the algorithm itself. Despite its advantages, the Louvain method is known to be non-deterministic and not robust against massive noise in the input data. Although in our experiments the Louvain method yielded stable partitions, this stability is not guaranteed when dealing with other types of networks. In order to overcome the stability problem, alternative community detection algorithms should be considered. A good alternative is the Markov Clustering algorithm due to its robust behaviour and tolerance to noise. Another option is the stabilised version of the Louvain method developed by Aynaud and Guillaume (2010).

Comparison between the landmark and the sliding window models: the analysis of the dynamics of the customer network using the sliding window approach allows us to identify the migration of customers between communities. This kind of business events remains potentially undetected when relying only on a landmark window approach since, in such a case, the customer would appear as belonging to a single larger community. We exemplify this type of event (*i.e.*, customer's migration) in Figure 3.9. In this figure we take a closer look at the dynamic customer network (sliding window approach) by focusing on the movement of a specific customer. The identified customer is initially a member of a community characterised by customers who buy a specific type of product. As time goes by, this customer starts buying other products that are associated to a different community. As a consequence, the customer begins to gradually approach

the second community and, as the old connection is forgotten, the customer moves to the only community it is now connected to. Due to the forgetting mechanism of the sliding window, the most up-to-date membership is highlighted thus enabling the detection of the customer's transition from one community (or profile) to the other. Having the knowledge of the underlying data and of the company's business model, we are able to identify that the purchasing of the second product is a natural action after the purchasing of the first one. However, there are other companies which also supply the latter and, if this transition could be predicted, the company could have sold both products as a package in the first instance. This type of proactive commercial actions would reduce the company's commercial risk and increase its sales volume.

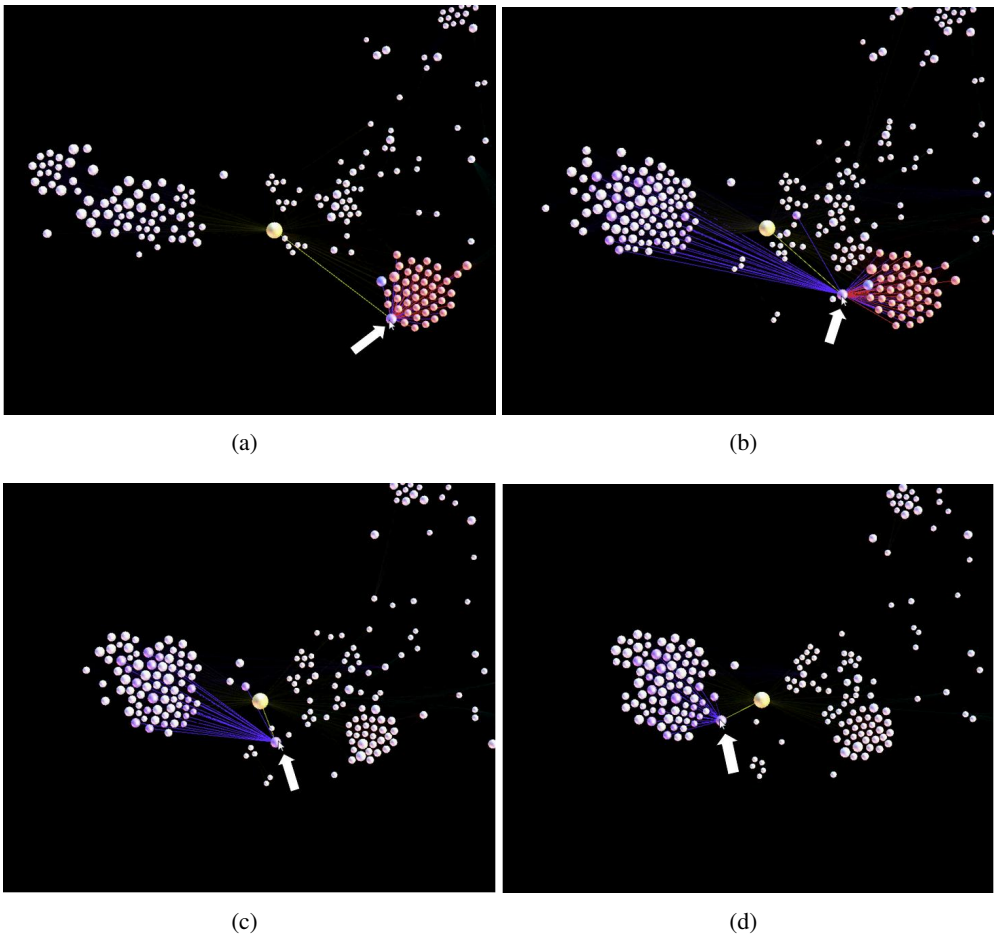


Figure 3.9: Illustration of the migration of a customer from one community to another. The customer's movement is marked by an arrow.

3.5.4 Additional Experiments

Sensitivity Analysis of the Matching Threshold of MECnet

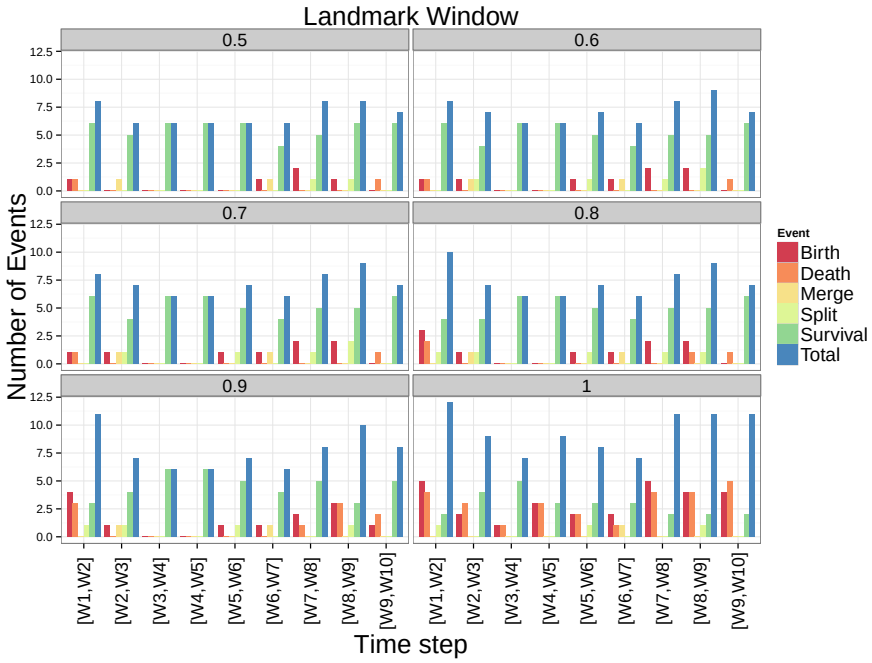
As mentioned in Section 3.4.2, the mapping process underlying MECnet relies on a user-defined threshold - the matching threshold τ-, which takes values from the interval $[0.5, 1]$. Here, we perform a sensitivity analysis of τ in order to assess the influence of choosing different values for the matching threshold in both the number and type of detected events. We run experiments with different values of τ, for both window models. The results are reported in Figure 3.10.

The analysis of the results for both window models suggests that higher values of τ increases the number of detected births and deaths, and decreases the number of detected survivals and merges. We also observe a rise in the total number of detected critical evolutionary events. This finding agrees with the intuition behind the concept of matching threshold. When τ increases, we are imposing a stricter condition on MECnet for detecting communities' matches, which naturally reflects in a more conservative matching behaviour. If we set $\tau = 1$, MECnet will only consider a match between two community instances $Com^{m}_{t_i}$ and $Com^{u}_{t_i+\Delta t}$ if all nodes pertaining to community $Com^{m}_{t_i}$ migrate to community $Com^{u}_{t_i+\Delta t}$. This implies that, what was once categorised as a single survival event, will now be likely replaced by two events: a death and a birth. As a consequence, the total number of critical evolutionary events increases. An analogous rationale applies to the merge events, since our definition requires that there are at least two communities' matches for a merge to be detected. So, if the match condition (τ) becomes more demanding, the detection of merges becomes less likely as well.

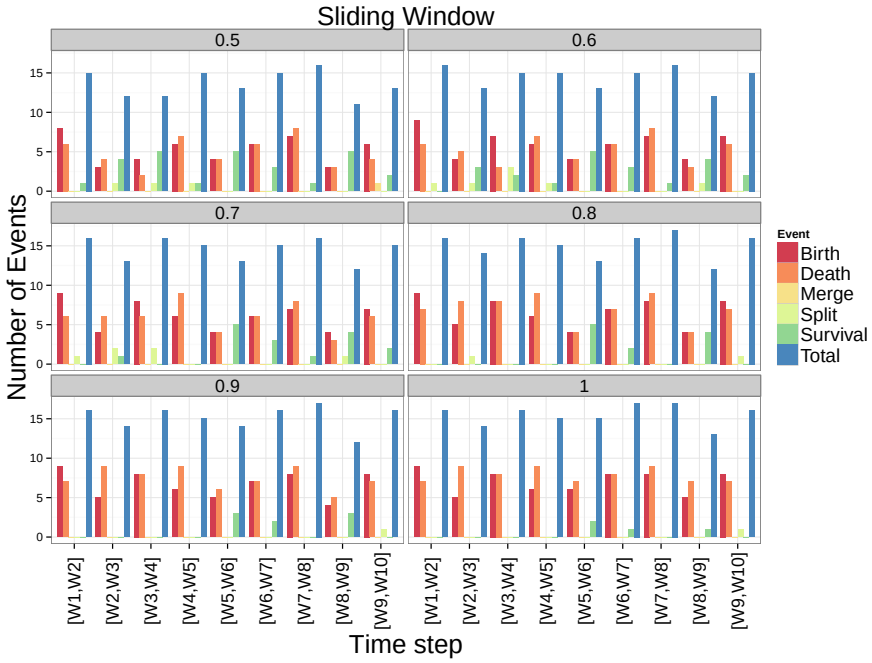
Based on this analysis, we conclude that the choice of the threshold τ influences both the number and the type of events detected by MECnet. Lower values of τ are more flexible, allowing the detection of more survivals, which reflects in longer communities' life-cycles. In turn, setting more demanding values for τ shortens the communities' life-cycles due to the detection of disruptive events, such as births and deaths. The survivals detected for strict values of τ are, however, indicative of highly stable communities (or customer profiles), whose detection might be useful for certain applications.

MECnet with Label Propagation Algorithm

In order to assess the influence of the choice of the static community detection method on the MECnet results, we conducted experiments using the LP algorithm. We applied the LP algorithm to each network snapshot of each window model. In order to ensure a fair comparison with the Louvain method, we selected the communities containing at least 75% of nodes and discarded the remaining ones. For the landmark windows, LP detected an average of 5 communities ($\sigma = 2$), a maximum of 10, and a minimum of 3 communities, depending on the considered time step. The average size of communities was 109 nodes, with the smallest group having 10 nodes, at time



(a) **Landmark window**



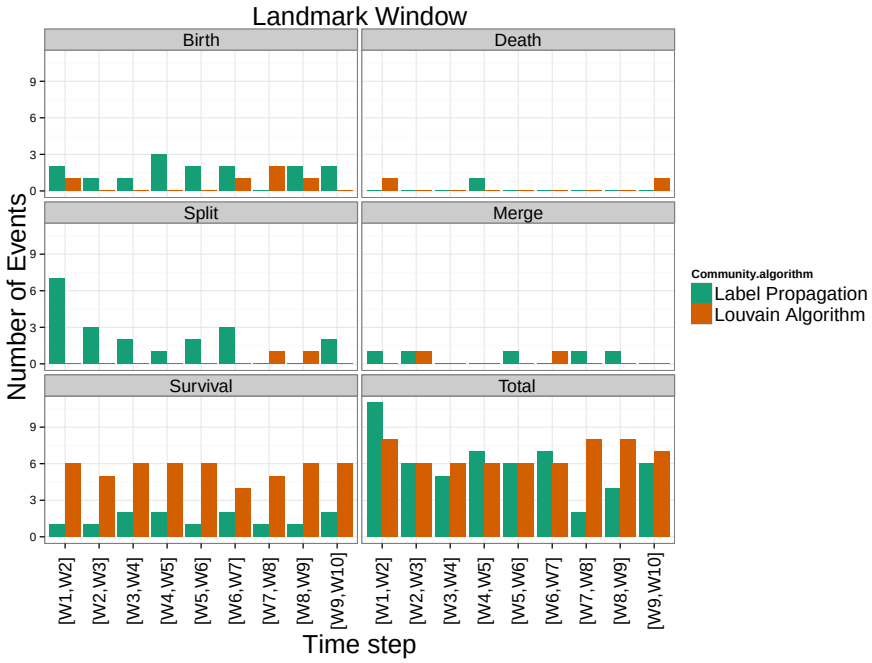
(b) **Sliding window**

Figure 3.10: Influence of the matching threshold of MECnet ($\tau \in [0.5, 1]$) on the number of critical evolutionary events detected (births, deaths, splits, merges, and survivals).

step W_1, and the largest group containing 381 nodes, at time step W_8. Regardless of the variability observed in both the number and size of communities, all community structures were meaningful, as indicated by the average modularity of 0.66 ($\sigma = 0.02$). We obtain slightly different results for the sliding windows. Using LP, we extracted an average of 10 communities ($\sigma = 2$), a maximum of 13, and a minimum of 6 communities. From the whole set of communities, the smallest one was found at time step W_5 and was comprised of only 6 nodes, whereas the largest community was detected at time step W_6 and contained 106 nodes. The average modularity was 0.58 ($\sigma = 0.09$). Although the modularity average value decreased, when compared to the landmark window scenario, it is still high and indicative of the existence of meaningful customer profiles. Comparing these results with the ones obtained by the Louvain method (Figure 3.5), we conclude that, although we do not observe a stark contrast in the values of these indicators for each algorithm, it is apparent a higher variability in the number of communities over time steps when using the LP algorithm. Besides, according to the modularity criterion, the quality of the network partitions is on average higher for the Louvain method.

Regarding community dynamics, we followed the same procedure used for the Louvain method. The number of critical evolutionary events returned by MECnet for each static community detection algorithm, and for each window model, are shown in Figure 3.11. Starting with the landmark window model (Figure 3.11-(a)), we observe that when using the community structures discovered by the LP algorithm the number of splits and births is much higher than the ones detected using the Louvain method. This increase in splits and births is compensated by a reduction in the number of survivals. The apparent inverse relationship between the splits/births and survivals might be a consequence of the higher variability on the number of communities detected by the LP algorithm. In fact, if the number of customer profiles is small in a given time interval, and large in the next time interval, the customer profiles either splitted into several profiles or new profiles were created. Concerning the sliding window (Figure 3.11-(b)), the total number of events detected by MECnet for the LP algorithm is consistently larger than the ones discovered for the Louvain method. By taking a closer look at the graphs, we can deduce that this large number of events is a reflection of the higher number of detected births, deaths, and splits. Despite this high dynamicity on the evolution of communities, which is captured by both community detection algorithm and can be explained by the forgetting mechanism employed by the sliding window model, LP appears to extract more unstable community structures than the Louvain method. Once again, a possible justification for this difference resides on the variability of both the number and the composition of LP's communities.

Based on the performed comparison, we conclude that the choice of the static community detection method influences the perceived dynamicity of communities, and the number and type of events captured by MECnet.



(a) **Landmark window**

(b) **Sliding window**

Figure 3.11: Influence of the static community detection method (Label Propagation algorithm and Louvain method) on the number of critical evolutionary events identified by MECnet.

Comparison of MECnet with the GED method

In this section, we compare the results returned by MECnet with the ones obtained using the Group Evolution Discovery (GED) method (Bródka et al., 2013). Similarly to MECnet, GED is an event-based framework for community evolution analysis in social networks that relies on a two-stage approach. During the first stage, communities (or groups) are discovered at each snapshot of a given dynamic network by resorting to an arbitrary static community detection algorithm. Then, the relevance of nodes pertaining to each community is computed using any node importance measure (*e.g.*, degree centrality, betweenness centrality, PageRank, social position). In the second stage, the *inclusion* measure is calculated for each pair of community instances found at consecutive time steps. *Inclusion* takes into account not only the size of the community membership intersection between pairs of community instances, but also the quality of nodes migrating from one community instance to another. Based on the values of the inclusion measure and the values of three user-defined parameters (α, β, and forming/dissolving threshold, also referred to as fd), the method looks for the presence of seven types of critical events: continuing (equivalent to MECnet's survival), splitting (equivalent to MECnet's split), merging (equivalent to MECnet's merge), forming (equivalent to MECnet's birth), dissolving (equivalent to MECnet's death), growing, and shrinking. The two additional events considered in the GED method (growing and shrinking) are internal events, in the sense that they capture changes concerning the contents of each community. In MECnet we focus only on external events, which relate to changes occurring in the whole community structure. Nevertheless, MECnet can be easily extended to accommodate these kind of internal events, by keeping track of the cardinality of communities.

In order to ensure a fair comparison between MECnet and GED, we considered the same set of communities returned by the Louvain method, which were used as input for MECnet. Regarding the node importance measure required by GED, we selected the PageRank (Brin and Page, 1998; Page et al., 1999) instead of other centrality measures, such as degree or betweenness centralities, due to its ability to take into account both the quantity and quality of the nodes' connections within the network. PageRank was computed for the whole set of nodes that were active at each time step. We run the original implementation of GED method available in Piotr Bródka's web page [1]. For our experiments, we used the default values of GED's parameters: $\alpha = 50\%$, $\beta = 50\%$, and $fd = 10\%$. Since the parameter α is the GED's equivalent to MECnet's τ, the chosen α value (*i.e.*, $\alpha = 50\%$) is consistent with the value chosen for τ in our case study's experiments (*i.e.*, $\tau = 0.5$). Such consistency guarantees a similar level of flexibility in both methods when looking for communities' matches.

We run the GED method for each window model. A comparison of the number and type of events detected by MECnet and the GED method is presented in Table 3.5. For the land-

[1] `http://www.ii.pwr.wroc.pl/~brodka/ged.php`

Table 3.5: Comparison of the number and type of events detected by two event-based frameworks for community evolution analysis: the MECnet and GED frameworks, for all time step intervals of the dynamic customer network. The proportion of external events (*i.e.*, growing and shrinking events were excluded from the computation) with relation to the total number of external events, for both methods, is shown inside brackets.

Landmark Window								
Method/ Event Type	**Birth/ Forming**	**Death/ Dissolving**	**Split/ Splitting**	**Merge/ Merging**	**Survival/ Continuing**	**Growing**	**Shrinking**	**Total**
MECnet	72 (17.6%)	46 (11.2%)	24 (5.9%)	11 (2.7%)	256 (62.6%)	-	-	409
GED method	2088 (36%)	2376 (40.9%)	1344 (23.1%)	0	0	36	90	5934
Sliding Window								
Method/ Event Type	**Birth/ Forming**	**Death/ Dissolving**	**Split/ Splitting**	**Merge/ Merging**	**Survival/ Continuing**	**Growing**	**Shrinking**	**Total**
MECnet	338 (42.6%)	347 (43.8%)	20 (2.5%)	1 (0.1%)	87 (11%)	-	-	793
GED method	2736 (39.5%)	2880 (41.6%)	1241 (17.9%)	72 (1%)	0	93	78	7100

mark windows, the GED method detected a total of 5934 events for all time step intervals ($[W_1, W_2], ..., [W_9, W_{10}]$). MECnet extracted a total of 409 events for the same data. The stark contrast between the total number of events of MECnet and the ones returned by GED are explained by GED's assumption that a single community instance found at time step t_i might be involved in several events with other community instances at a later time step $t_{i+\Delta t}$. For instance, while MECnet categorises a split of one community into three communities as a single event (*i.e.*, one split), the GED method considers this occurrence as three splitting events. From the 5934 events discovered by GED, 2376 were dissolving, 2088 were forming, 1344 were splitting, 36 were growing, and 90 were shrinking. Contrary to MECnet, no continuing events were detected by the GED method. This might be influenced by the fact that, while MECnet only takes into account the quantity of nodes shared by two community instances, the mapping process of the GED method also considers the position and importance of the nodes belonging to the community instances, as measured by PageRank. A possible explanation to this is the fact that GED method considers as dissolving events communities that remain inactive over several time frames, which is likely to happen in our data due to the company's business nature. When these communities become active again, the GED method categorises their evolution as a forming event. Such cases are considered as survivals when using MECnet. Regarding the overlapping sliding windows, the application of the GED method for all time step intervals returns a total of 7100 events. This raise in the number of total events, with respect to the landmark windows, somehow reflects the higher dynamicity of the networks obtained by the sliding window approach. This increase in the overall number of extracted events for the sliding window is consistent in both MECnet and GED. These 7100 events discovered by GED

are broken down into 2880 dissolving, 2736 forming, 1241 splitting, 72 merging, 93 growing, and 78 shrinking events. As expected, the number of forming/dissolving events is much larger for the sliding window than for the landmark window. This finding agrees with the results of MECnet for the sliding window, as suggested by the identical proportions of births/deaths (forming/dissolving in the terminology of GED) obtained by both methods (please see Table 3.5).

Computation times of both methods for such small networks are not relevant, given that the experiments took only a few seconds.

3.6 Conclusions and Future Work

We introduce an application-driven methodology to analyse community structure dynamics in evolving customer networks, based on two window models: a landmark and a sliding window. The methodology encompasses an event-based framework, dubbed MECnet, to identify and monitor dynamic communities. Although the proposed MECnet framework is general and has several applications, the introduced methodology is application-driven and its development was motivated by the need of companies to tap the potential of customer data for marketing purposes. In this chapter, we present the results of our empirical analysis on the time-stamped customer base data, spanning one year, of one of the largest Portuguese companies on the field of electricity. This dynamic network uncovers the similarities of purchasing behaviour among the company's customers over time. The study of several time series of SNA measures, and the application of MECnet using two different time window models, allowed us to identify the evolutionary profile of groups of customers and grasp insights into the customer base. The average modularity value of the community structures identified by the Louvain method is higher than 0.6 for both window models suggesting the existence of well defined and meaningful purchasing profiles of customers. The results also indicate that both window models deliver different but complementary types of knowledge regarding the network dynamics, which is consistent with the literature results. While the landmark window considers all the historical data, the overlapping sliding window employs a catastrophic forgetting of older data, focusing only on the most recent past. Given these distinct perspectives, the sliding window approach proves to be more suitable for, *e.g.*, detecting acute changes in customers' purchasing behaviour, whereas landmark windows are more appropriate to identify persistent, enduring, and stable evolutionary customer profiles. Thus, in cases where the company does not have any information regarding the dynamicity of the customer network, we suggest using both types of windows in the analysis of evolutionary customer profiles. These conclusions were drawn from the analysis of the number of critical evolutionary events and the survival ratio derived from the application of MECnet. The presence of persistent and stable profiles in the dynamic customer network when using a landmark window model were revealed by the high values of the survival ratio and the high number of MECnet's community survivals. On the other

hand, the dynamicity of the network when using the overlapping sliding window model manifested itself by low values of the survival ratio and by a large number of community births followed by community deaths.

From a business point of view, the introduced methodology can be useful to Marketing teams since it allows the detection, monitoring, and analysis of the dynamics of customer profiles at different levels of temporal resolution. These profiles are discovered based on the similarity of customers' behaviour, thus providing explicit information about customers' needs and preferences, and how they change over time. The temporal analysis of such purchasing profiles may help companies (i) acquire a deeper knowledge about their customers' preferences, (ii) predict the type of product the customer will buy next, or (iii) support the development of recommender systems. From the deployment point of view, our methodology has several desirable features: (i) it relies on user-defined thresholds and (ii) it is incremental. MECnet thresholds (τ and λ) introduce flexibility in the framework, allowing managers to flexibly adjust the scope of monitoring and incorporate their business knowledge in the analysis. The incremental nature of the method is also advantageous from the resources perspective. Since the analysis is based on snapshots of the network, both SNA measures and the MECnet model can be updated as new network snapshots are available, ensuring timely results. The update of results can be done by first computing the SNA measures and detecting communities in the newly arrived network snapshot, and then performing the comparison of these results with former values of the measures and the community instances discovered in the previous time step. If the analysis is performed under a sliding window model, the analysis is more efficient in terms of memory size and processing time since the newly arrived network snapshot is expected to contain less data than the corresponding landmark window snapshot.

Despite these advantages, the research work presented here has a few limitations, which are mostly related to the assumptions made. First, MECnet assumes communities are exclusive or crisp, *i.e.*, disjoints sets of entities. This implies that each customer is affiliated with exactly one community at a time, though its membership may shift over time. This assumption does not always hold, since communities may overlap (*e.g.*, in a customer network, if a customer purchased two types of products, and these two types are not usually purchased by other customers, its purchasing behaviour can be associated with two distinct communities or profiles). Secondly, the dynamic community mining frameworks proposed so far lack strategies to understand the reasons behind the detected events. These reasons can be investigated by inspecting the data *a posteriori* but are not considered in the framework itself. Thirdly, when determining the correspondences between community instances found at consecutive time steps using conditional probabilities, we assume that all community members have the same importance. However, communities may exhibit a core-periphery structure, where densely-connected core entities coexist with sparsely-connected peripheral entities (Rombach et al., 2014). It may be hypothesized that these core entities are those responsible for the survival of a community over time, whereas the peripheral entities have a volatile

behaviour in what regards community membership. Thus, under this scenario, core community entities play a more important role in the stability of a community and should be assigned higher weights in the matching procedure. The fourth identified limitation is related to the difficulty of choosing the most appropriate window length and step width. In this work, we have relied on the domain knowledge of the Business Intelligence analyst to determine interval lengths for snapshots but this strategy can be problematic in the absence of domain information, since an improper choice of these parameters has effect on the type of obtained results. Two other limitations are related to the number of window models tested, and the size of both the dynamic network and the time frame.

The above-mentioned limitations offer several directions for further research. The first limitation can be addressed by applying, to each network snapshot, static community detection methods that are able to discover overlapping communities (*e.g.*, using the CPM method of Palla et al. (2005) and Derényi et al. (2005)). This would imply a relaxation of the assumption of crisp communities underpinning MECnet and, consequently, it would require an adaptation of the mapping procedure so as to handle the more complex output generated by these methods. The second limitation could be overcomed by introducing additional steps aimed at expressing the detected changes in a comprehensible way. We tackle this problem in the next chapter, where we propose a trajectory-based visualisation approach that promotes the understanding of the reasons behind the changes in dynamic communities, in terms of structural and relational information extracted from the network. Concerning the third limitation, the core-periphery structure is more likely to be observed in typical social networks (*e.g.*, scientific co-authorship networks) and not always emerges in other types of networks, such as the one we studied. In the customer network, the adopted definition of relationship among customers generates communities with a more homogeneous degree distribution than communities displaying a core-periphery structure. Therefore, it is reasonable to assume that all community members have similar importance. To handle the fourth limitation, methods such as time series segmentation and smoothing can be used to determine appropriate window lengths for the network snapshots. The last two mentioned limitations can be circumvented by performing experiments using alternative time window models (*e.g.*, accumulated time windows with fading links, tilted time windows) and a few larger datasets, with longer time frames. These experiments are relevant for empirically assessing the feasibility, performance, and suitability of our MECnet for dealing with very large networks.

As future work, it would also be interesting to pursue the following ideas. Companies have access to rich information about their customers, which can be better harnessed for finding meaningful customer communities. Information about customers' individual attributes, such as gender, age, and spending patterns, could be incorporated in the community detection process in order to obtain richer and more complete profiles. An idea would be to explore community discovery methods that combine linkage structure with node content (*e.g.*, the NEIWalk method of Wang et al. (2014)) for identifying evolutionary customer profiles. From the application perspective, it would also

be interesting to apply the proposed methodology to data from the Business-to-Consumer (B2C) market, such as those collected by the retail sector, because the characteristics and dynamicity of the corresponding customer networks are expected to be different. Another possible extension of this work is the analysis of the product networks associated with the customer networks. The exploration of the dynamics of a product network is interesting as it uncovers sets of frequently bought products (the communities) and highlights temporal changes in these sets.

Chapter 4

Visualisation of Dynamic Networks using Spatio-Temporal Trajectories

Visualisation of static social networks is a mature research field in information visualisation. Conventional approaches rely on node-link diagrams, or graphs, and matrix-based representations, such as adjacency matrices. However, the widening availability of longitudinal network data has spurred interest in visualisation approaches that go beyond the static node-link or matrix-based representations of a network. In dynamic settings, the focus is on the exploration of network dynamics at different levels of analysis (*e.g.*, node, communities). Yet, the development of visualisations that are able to offer insight into different types of changes occurring in an evolving network is a challenging task. In such settings, extensions of the traditional node-link and matrix-based representations can prove to be limited. This work attempts to overcome these limitations by proposing a novel methodology for modelling and visualising the evolution of dynamic social networks, at both the node-level and the community-level, based on the concept of spatio-temporal trajectory. We resort to a three-order tensor to represent structurally-derived statistics of an evolving network and we further decompose it using a Tucker3 model. The two most representative components of this model, associated with the structural statistics dimension, define the 2D common space where the trajectories of social entities (*i.e.*, nodes or communities) are projected. To illustrate the proposed methodology we conduct a case study using a small set of temporal self-reported friendship networks.

4.1 Introduction

The network perspective views complex systems as sets of interrelated entities. This perspective is anchored on the assumption that analysing the relationships among entities can reveal valuable information which cannot be accessed by merely analysing the entities' individual attributes. The

information associated with the system's relational structure can be derived by performing *network analysis*. Network analysis provides a formal and powerful mechanism for modelling, representing, describing, and measuring such relational structures and, given the ubiquity of networks in our world, it has been embraced by virtually all of the sciences, from social psychology to physics. One of the most widely studied type of networks is the *social network*. A social network details the relationships, or interactions, among social entities, which can be individuals or virtually any aggregation of individuals, such as groups, communities or organisations. The underlying structure of such networks is the object of study of SNA. Information visualisation has a long history in SNA and has been used to support the analysis of networks since the seminal work of Jacob L. Moreno (1934) in the 1930s. Moreno (1934) was the founder of sociometry, a field of social psychology drawing on the ideas of the *gestalt theory* and concerned with the analysis of small group dynamics and their impact on personal choices. One of his greatest innovations was the development of the *sociogram*, a systematic method to represent social structures as diagrams comprised of points and lines. Such diagrams revealed to be very useful in uncovering the hidden structures of groups, by means of the identification of stars, alliances, and subgroups. Since then social scientists have been using visual representations both to explore the structure and patterns of social networks and to communicate their findings. There are two conventional approaches for visually representing static social networks (Henry and Fekete, 2006; Beck et al., 2014): node-link diagrams (or sociograms), that provide a representation of the network topology by representing nodes as points and links between them as lines (this is equivalent to the graph model); and matrix-based representations, such as adjacency matrices, where in the simplest case coloured cells indicate whether two nodes are connected by an edge. The node-link diagrams provide geometric representations for relatively small networks that are intuitively understood by the users (Lee et al., 2006), thus remaining a powerful means of communication. There are several graph drawing algorithms for producing node-link diagrams, such as the following well known force-directed layout algorithms: the Kamada-Kawai algorithm (Kamada and Kawai, 1989) and the Fruchterman-Reingold algorithm (Fruchterman and Reingold, 1991) (for a comprehensive survey of graph layout techniques, please see Gibson et al. (2013)). When dealing with large networks (*e.g.*, the WWW graph), these traditional graph drawing algorithms are unable to produce readable node-link visualisations. In such cases, a common strategy is to visualise the network at increasing levels of abstraction, such as at the group- or community-level (please see Brockenauer and Cornelsen (2001) for a survey of graph drawing algorithms for the visualisation of flat or hierarchical disjoint groups in graphs). Regarding adjacency matrices, these have the advantage of being almost always readable, even for large and dense graphs, and thus support exploration throughout the analysis process (Ghoniem et al., 2005). Besides, they are able to uncover high-level structures by finding good permutations of their rows and columns. This strategy is employed by the *blockmodelling approach* to find subsets of nodes in the network, the so-called *blocks*, that have

similar profiles of connections, *i.e.*, the nodes in each set are structurally equivalent (Lorrain and White, 1971) or regularly equivalent (White and Reitz, 1983).

Although these conventional approaches have some advantages, they were not originally tailored to represent networks that evolve over time through the addition and removal of nodes/links, and so they prove limited for the visualisation, analysis, and understanding of dynamic networks. The introduction of the temporal dimension in the visual representation poses additional challenges and increases the complexity of the solution (Brandes et al., 2012). In such settings, the extension of traditional node-link representations to dynamic networks can prove to be limited due to, for instance, the difficulty of generating discernible visualisations where nodes and edges encode temporal statistics or attributes, and also due to the inherent difficulty of humans in making sense of the evolution of a network over long temporal spans based on animations (Yi et al., 2010). Alternative methods, such as matrix-based representations, are known to be less intuitive, computationally inefficient for representing large-scale networks, and may fail in tasks involving path finding (Ghoniem et al., 2005). In this work, we deviate from these conventional approaches and propose a novel visualisation based on spatio-temporal trajectories that provides a static but informative view of the evolution of the dynamic network at multiple levels of analysis.

In order to produce an efficient and interpretable network visualisation it is desirable to ensure its conformity to a set of graphical principles. Drawing on the ideas of Tufte and Graves-Morris (1983), Brandes et al. (2006) argue that a good graphical representation of networks should provide the viewer with the greatest number of ideas in the shortest time and space, by communicating clearly, accurately, and efficiently the most interesting and relevant information contained in the raw relational data. Due to the increased complexity of dynamic networks as compared to static networks, the adherence to such principles becomes even more critical. By focusing on the essence (or patterns) of the network instead of simply reporting the available data, it is expected to produce a visualisation approach that brings the viewer closer to knowledge. In the context of dynamic networks, two major issues are the identification of the moment when changes occur and the understanding of the nature of those changes with respect to interpretable network constructs, such as centrality and reachability. In this work, we address these issues by proposing an efficient visualisation methodology that focuses on capturing and semantically understanding the changes occurring in the structure of a dynamic network at multiple levels of analysis. The efficiency of the proposed visualisation is achieved by summarising the network data and displaying only the most important information gathered from the network structure. By employing this strategy, we are able to produce a merged view of the entire historical evolution in an informative static bidimensional space.

Our methodology enables the visualisation of the structure of evolving social networks at two different levels of analysis, the node-level and the community-level, using as a basis the concept of *spatio-temporal trajectory*. In order to meet the above-mentioned graphical principles (Tufte

and Graves-Morris, 1983; Brandes et al., 2006), we propose the generation and interpretation of trajectories of social entities in a common bidimensional space (2D), which accounts for most variation in the original data, based on the output of a Tucker3 model with orthogonality constraints. These social entities can be nodes or groups of nodes that interact more often with each other (*i.e.*, communities). Similarly to Ahn et al. (2011), throughout this chapter we use the term *social entities* to refer to the objects analysts are interested in (*e.g.*, nodes or communities). By allowing the visualisation of different types of social entities, our approach is able to support the analysis of the temporal activity of networks at multiple levels. This becomes particularly useful when analysing large dynamic networks, for which prohibitive amounts of time are required to analyse each individual node. In such cases, it may be more efficient and interesting for the analyst to examine and get a meaningful overview of the dynamics of the network at higher-levels of abstraction (*e.g*, at the community-level).

As an example, consider a group of students entering university and facing a new social environment. While a few students may find acquaintances and friends from high-school attending the same university, it is likely that most of them do not know their classmates. When facing this kind of social contexts, individuals tend to invest more effort in social communication so as to develop friendly relationships with their peers. This socialisation process is expected to be more intense during the first weeks of university, which typically correspond to periods of social exploration and adaptation. However, after some time the friendship groups start to settle down and the intensity of students' socialisation tends to stabilise. Since the relationships established among university freshmen over time can be mapped as an evolving social network, the socialisation phenomenon can be investigated by means of the proposed methodology. Our visualisation would not only provide an overview of the general evolution of the social behaviour of the whole set of students, but would also allow to delve into questions concerned with the social dynamics of each student (*e.g.*, changes in popularity and the moments when they occurred) or with the evolution of each group of students (*e.g.*, identification of temporally cohesive groups and/or highly dynamic groups, and examination of the corresponding reasons).

The main contribution of this work is thus a novel methodology for modelling and visualising evolving networks (comprised of a finite number of snapshots) that relies on displaying spatio-temporal trajectories of nodes or communities in a common, compact, and interpretable 2D space. These spatio-temporal trajectories can be viewed as summaries of the temporal evolution of the dynamic network in a static 2D representation. While the proposed visualisation focuses on observing trajectories of single nodes or communities, some of the measures used to describe them might depend on the global network structure (*e.g.*, betweenness centrality, closeness centrality). In such cases, our approach is able to capture latent network properties in the 2D visualisation. This way, the resulting trajectories encapsulate both temporal and relational aspects of the network dynamics by indicating when a global or local change occurred and the meaning of such change.

The use of the Tucker3 model to map the information into a common 2D space guarantees that this space describes the most important information contained in the original three-way data while preserving the temporal interdependencies. On the other hand, since each axis of the 2D space corresponds to a component of the Tucker3 model, this space is interpretable in terms of the network measures used to describe the social entities comprising the network.

The remainder of this chapter is organised as follows. Section 4.2 provides the background on tensor algebra and introduces the foundations of the Tucker decomposition. Section 4.3 discusses related work. In Section 4.4, the problem of modelling and representing dynamic social networks is addressed and our approach to visualise their evolution using spatio-temporal trajectories is introduced. In Section 4.5 a case study using temporal friendship networks is presented, aiming to show how the methodology can be used in practice. Section 4.6 concludes the chapter and discusses possible avenues for future research.

4.2 Background

In this section, we introduce some preliminary concepts, terminology, and notation of tensor algebra, as well as the foundations of the Tucker3 model, which will be useful to understand the proposed methodology.

4.2.1 Tensors

The Concept of Tensor

Traditional data analysis techniques, such as linear regression, Principal Component Analysis (PCA), and clustering, are meant to extract relevant knowledge from two-way (also known as two-mode or two-order) data. Two-way data are represented through matrices, where rows typically correspond to the individuals or objects, and columns to the variables. Although two-order data constitute the basis of numerous and interesting analyses and are suitable for several applications, the variables are usually measured on a specific moment in time and, thus, only provide a static view of the studied phenomenon. Since many phenomena are inherently multidimensional and dynamic, in several settings one should adopt data representation schemes able to model simultaneously all dimensions, including the temporal one. In such cases, high-order tensors (also known as hypermatrices, multiway models, multiway arrays, or multidimensional arrays) appear as more natural and appropriate data representations than matrices, since they are able to explicitly model a higher number of dimensions (*e.g.*, objects, variables, and time) without collapsing the data and, therefore, without losing information about their mutual dependencies.

One of the desirable features of modelling data comprised of more than two modes as a high-order tensor, and explore techniques especially devised to deal with these data structures, is the

possibility of preserving all mutual dependencies established between the dimensions of different modes.

Tensor Notation

Regarding notation, we follow the typical conventions and use the standardised notation and terminology for multiway analysis as proposed by Kiers (2000). As previously mentioned, a tensor is a N-way data array, where N is the order (also referred to as ways) of the tensor. High-order tensors, denoted by calligraphic letter χ, are generalisations of scalars (order 0), vectors (order 1), and matrices (order 2), to three or higher orders. For instance, a three-order tensor ($N=3$) encapsulates three modes: the row-entities mode (mode A), the column-entities mode (mode B), and the fiber-entities mode (mode C). We will then use the term *mode* to refer to a set of entities. The element (i, j, k) of a three-order tensor χ is denoted by x_{ijk}, where index i (j and k, respectively) refers to the entities of mode A (mode B and mode C, respectively). Indexes typically range from 1 to their capital version: $i = 1, ..., I$, $j = 1, ..., J$ and $k = 1, ..., K$. For instance, in a longitudinal study where the characteristics of the same set of individuals are measured at different moments in time, x_{ijk} gives the score of individual i on variable j at time point k.

In the scope of our methodology, we will focus only on three-order tensors, although our methodology can also be extended to higher-order tensor representations.

4.2.2 Tucker3 Model

Three-way methods are unsupervised multivariate data analysis tools that compress simultaneous variation of combinations of variables and entities (Smilde, 1992), being often used for compression, decomposition, and denoising of data. These methods are useful for the exploratory and descriptive analysis of three-way arrays since they are able to summarise the information contained in these arrays, both in terms of main effects and two and three-way interactions, by means of a few number of components. Besides, these methods generate the so-called *core tensor* (or core array), which is a highly informative structure that captures and describes the relations and three-way interactions between the different modes of data, in terms of their summarised entities (Kiers and Van Mechelen, 2001). A major benefit of performing dimensionality reduction of three-order tensors is the possibility of plotting the most significant structure of the three-way data in a 2D or 3D space, by choosing as axes the components explaining the greatest variation in the data (Skillicorn et al., 2014). These low-dimensional spaces allow the visual inspection of the most salient patterns in the data. For a thorough review on unsupervised multiway analysis, please refer to Acar and Yener (2009).

The most widely known three-way methods are the CANDECOMP/PARAFAC (CP) (Carroll and Chang, 1970; Harshman, 1970) and the Tucker decomposition (Tucker, 1963, 1966). In this

work, we resort to the Tucker decomposition since it is more flexible, easier to interpret (in the sense that the solution can be rotated), and has less constraints than the CP. In fact, the CP model can be seen as a constrained variant of the Tucker decomposition, where the core tensor is super-diagonal and the cross-relations between components are eliminated.

Tucker (1963) introduced the tensor decomposition, which inherits his name, in 1963. Refinements of this model were then performed by Levin (1965) and Tucker (1966). Kroonenberg and De Leeuw (1980) further elaborated the three-way version of the model and named it *three-mode principal component analysis*. An in-depth study of Tucker decomposition was later undertaken by De Lathauwer et al. (2000), who coined its orthogonality-constrained version as High-Order Singular Value Decomposition (HOSVD). We briefly introduce the foundations of this decomposition following closely the definitions provided by Tucker (1966), Kroonenberg and De Leeuw (1980), De Lathauwer et al. (2000), Skillicorn (2007), and Kolda and Bade (2009).

The Tucker decomposition can be thought as a form of higher-order Principal Component Analysis or as a multilinear generalisation of the Singular Value Decomposition (SVD). The three-way version of this decomposition is known as the Tucker3 model. The term derives from the fact that the reduction of data is performed in all three modes of a three-order tensor (*i.e.*, the row-entities mode A, the column-entities mode B, and the fiber-entities mode C). The Tucker3 model expresses the three-order tensor χ as a product of component matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, and a small core tensor $\mathscr{G}$, in a way that reveals χ's internal structure. The general Tucker3 model can be written elementwise as:

$$\hat{x}_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr}, \tag{4.1}$$

where $i = 1,...,I$, $j = 1,...,J$, and $k = 1,...,K$. Here, the coefficients a_{ip}, b_{jq}, and c_{kr} represent the entries of orthonormal matrices, also referred to as component matrices $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$. These matrices can be thought of as the principal components in each mode. In turn, the coefficient g_{pqr} represents the entry of the so-called core tensor $\mathscr{G} \in \mathbb{R}^{P \times Q \times R}$. The number of entities in each mode are represented by letters I, J, and K. The number of components, or levels (*i.e.*, number of columns of the matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$) in the first, second, and third mode of the tensor are represented by letters P, Q, and R (usually, $P < I$, $Q < J$, and $R < K$), respectively. The component matrices summarise the entities of each mode into a few number of components and are able to reconstruct the most important patterns contained in data, whereas the core tensor can be seen as an idealised summary or strongly reduced version of the full data array χ, given in terms of the summarised components for the three different modes. The core tensor $\mathscr{G}$ can be interpreted as a compressed version of the original tensor χ if P, Q, and R are smaller than I, J, and K. Tucker (1966) suggested interpreting the core tensor as describing the latent structure in data, since it has information about the level of interaction between the different components, and the

component matrices as mixing this structure to give the observed data. The core tensor values can also be interpreted as a generalisation of the eigenvalues, or of the singular values of the SVD, and they constitute a further partitioning of the "explained" variation as is indicated by the eigenvalues of the standard PCA. In fact, when the component matrices are orthogonal, the square of each entry of $\mathscr{G}$ is proportional to the amount of variance explained by the corresponding combination of components, in terms of total fit. Besides, the sign and magnitude of the core tensor values provide clues on how the various components relate to each other, with larger absolute values meaning a larger contribution to the final reconstruction of the input tensor. Core tensor values can also be interpreted as weights giving the relative importance of each combination of components. Matrices **A**, **B**, and **C** are usually constrained to be columnwise orthogonal. The orthogonality is desirable since it facilitates the analysis and hastens the computation of the decomposition, without compromising the optimal model fit to be achieved.

The basic idea of the tensor decomposition proposed by Tucker (1966) is thus to find those components that best capture the variation in each mode. Or, in other words, the goal of Tucker's method is to find a set of matrices **A**, **B**, and **C**, and a small tensor $\mathscr{G}$ that, in general, have less dimensionality than the original tensor, but are able to reconstruct the most important information, or patterns, contained in data, as depicted in Figure 4.1.

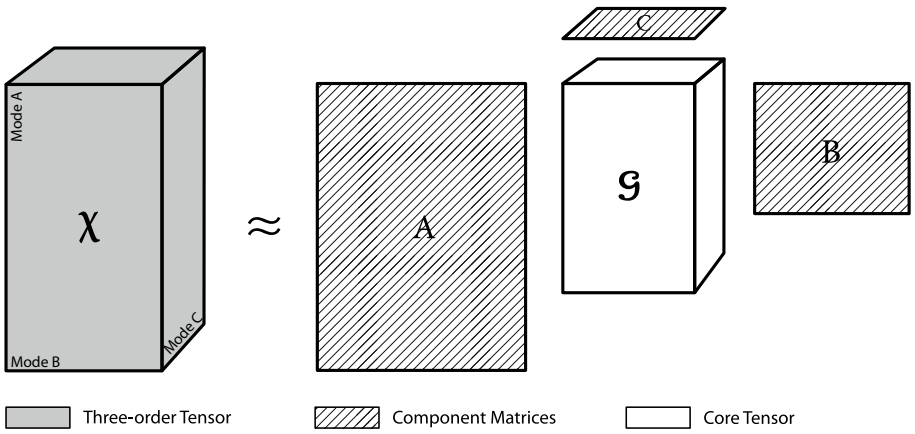


Figure 4.1: The Tucker3 decomposition of a three-order tensor.

The problem of decomposing a tensor, which can be translated into finding the best estimation of the model presented in Equation (4.1), can be reduced to a straightforward optimisation problem (*i.e.*, minimization of the difference between the original and the estimated model), which is usually solved using an Alternating Least Squares (ALS) approach. More detailed information about the ALS algorithm can be found in Kroonenberg (1983).

4.3 Related Work

Visualisation of evolving data fosters the understanding of the nature of temporal changes and it is essential for quickly and clearly conveying the results of the analysis (Moody et al., 2005; Bender-deMoll and McFarland, 2006; Brandes et al., 2006). The importance of visualisation has long been recognised by many researchers and a plethora of literature on visualisation of temporal data has been published in the past decades (a comprehensive survey can be found in Aigner et al. (2007)). More recently, the problem of dynamic graph visualisation gained momentum and established itself as an important research topic in the fields of information visualisation and graph drawing. The majority of solutions proposed so far are basically extensions of the prevailing approach (the static node-link representation) to dynamic networks (for a survey, please see Archambault et al. (2014)). These solutions focus on the creation of smooth animations between changing sequences of graphs. The main challenge in this kind of approaches is to balance the layout stability with the layout quality among frames, in order to guarantee the creation of readable animations able to preserve the viewer's mental map (Archambault et al., 2014; Beck et al., 2014). However, these solutions only work well on small and sparse networks. Approaches deviating from this research line include dynamic maps, timeline-based representations, and link reduction algorithms. The former are inspired by the geographic map metaphor to generate animations of map-like representations of the dynamic network. These representations encode additional node attribute information using contours and heat-map overlays (Mashima et al., 2012). Timeline-based approaches depict the temporal dynamics of networks by providing the complete sequence of graphs in a single chronologically ordered diagram (Ogawa and Ma, 2010; Reda et al., 2011; Tanahashi and Ma, 2012; van den Elzen et al., 2014). In this research line, a visualisation approach that is somewhat related to ours is the one by Reda et al. (2011), who introduce a *community structure timeline* that allows the inspection of the dynamic network at the community-level based on the community affiliation shifts of individuals. The proposed visualisation relies on a timeline chart that depicts the temporal evolution of communities by displaying the passage of time in the x-axis and the identified communities in the y-axis. Communities are defined as sets of individuals. A single thread/line is used to visually plot the community affiliation of an individual across the time axis. Individuals that belong to the same community at a given time step form a bundle of threads. This bundle denotes the existence of a community. These individuals are allowed to shift their community affiliation as the network evolves and this is captured in the visualisation by crooked threads/lines. The visualisation also incorporates information about nodes' attributes, by colour coding the individuals' threads based on the corresponding attribute values. Hence, the resulting community structure timeline enables the viewer to correlate patterns of affiliation shifts with these attributes, offering insights into the reasons behind individual changes of community membership. Hybrid approaches combining the animation of node-link diagrams with static timeline-based representations were

also proposed (see, for instance, the work of Sallaberry et al. (2013)). Link reduction algorithms, such as minimum spanning trees and Pathfinder networks, were used for visualising the evolution of scientific co-citation networks in Chen and Morris (2003). The visualisations derived by each technique were qualitatively evaluated based on two criteria, namely, the topological properties and the dynamical properties. Based on this evaluation, Chen and Morris (2003) concluded that Pathfinder networks are more intuitive and able to better preserve the evolution of the network when compared to the minimum spanning trees. Similarly to node-link representations, conventional matrix-based representations were also extended to dynamic settings. According to Beck et al. (2014), two types of matrix-based approaches can be identified: the intra-cell timelines, where the changes associated with the dynamic edges are encoded in the corresponding cells of the adjacency matrix by using timelines (*e.g.*, gestaltlines, bar graphs) (Brandes and Nick, 2011; Burch et al., 2013); and the layered matrices, where the adjacency matrices associated with each time step of the evolving network are sequentially displayed side-by-side using a small-multiples representation; or stacked into a 3D representation (Bach et al., 2014); or even radially distributed in a single diagram (Vehlow et al., 2013). However, these solutions are not always intuitive and easily interpreted by an inexperienced analyst, due to their visual complexity, and may fail in conveying the most relevant or interesting patterns in the network evolution, due to their focus on network description.

Regarding the visualisation of multidimensional data, traditional approaches include projections of data entities in low-dimensional spaces, that preserve most of the original data variation or the distances among data points. Examples of such representations for two-way data are the ones yielded by MultiDimensional Scaling (MDS) and PCA, where the entities are usually plotted in a 2D space. On the other hand, when dealing with tensorial representations of data, the analysis of their decompositions is commonly performed through the analysis of graphical representations of the component matrices of each mode, where the x-axis represents the entities of the corresponding mode and the y-axis represents the entries of the component matrix (see, for instance, Dunlavy et al. (2011) and Khayamian et al. (2012)). However, even though the generated visualisations are clear and able to represent the most important patterns found in data, information regarding, for instance, the evolution of the entities is somehow difficult to grasp without resorting to additional visual cues. We consider that by projecting spatio-temporal trajectories depicting the entity's dynamics in these low-dimensional spaces, a clearer and more efficient visualisation is obtained.

Nevertheless, to the best of our knowledge, no work to date has framed the problem of modelling and analysing dynamic networks as we did, nor proposed a visualisation approach based on spatio-temporal trajectories using Tucker spaces for this purpose. Yet, there are related work that resembles ours in one or two key steps of the methodology, although they have important differences. The first is the work of Bader et al. (2006). In this research, the authors identified patterns in the evolution of a directed email network by first modelling the time-varying relational data as a three-order tensor of weighted adjacency matrices (nodes $\times$ nodes $\times$ time), and then by applying a three-way

DEcomposition into DIrectional COMponents (DEDICOM) model. However, their analysis is solely based on the interpretation of the output of a tensor decomposition model (in this case, the three-way DEDICOM) and no visualisation is presented to help the analyst understand the results. Besides, they use the adjacency matrices to represent the networks, whereas we opt for a more informative data model embedded with structural information derived from the network. A similar modelling strategy is used by Dunlavy et al. (2011), who represent a dynamic social network as a temporally ordered sequence of adjacency matrices (node $\times$ node $\times$ time) and explore the outputs of matrix- and tensor-based decompositions for the task of link prediction. The research by Sun et al. (2009) is also related to the one addressed in our work in the sense they resort to three-order tensors to model social networks. These authors propose a hybrid two-step approach that summarises and extracts patterns from tensorial representations of large content-based social networks. On the first step, they perform a high-order dimensionality reduction using the Tucker decomposition on a three-order tensor detailing an email network (sender $\times$ recipient $\times$ keyword) and, in the second step, the dimensions of each one of the decomposed modes (sender, recipient, and keyword) are clustered. The extracted patterns are presented to the user by means of a hierarchical graph visualisation, which is build upon the results of the cluster analysis and allows exploration of the network at different levels of detail. Nevertheless, this approach was tailored for the analysis of static social networks and the possibility of extending it for handling dynamic social networks is not clear. In what regards the technical process of generating trajectories, the most related work to ours is the Structuration des Tableaux A Trois Indices de la Statistique (STATIS) method, proposed by Lavit et al. (1994). Similarly to the Tucker3 model, STATIS is a three-way method for exploratory data analysis. However, it differs from the Tucker3 model in the sense that it cannot be generalised to N-way data and performs separately the exploration of each tensor mode (instead of simultaneously, as in the Tucker3 model). One of the steps of this method consists in projecting the trajectories of entities over a given dimension (*e.g.*, time) in the so-called "compromise space". The main difference between ours and STATIS' trajectories lies in the space where we define them which, in our case, is based on the decomposition yielded by the Tucker3 model. Besides, in STATIS the interpretation of the trajectories is made with relation to the average trajectory of a fictitious entity, while in our approach the interpretation is based on the observable properties of the trajectories themselves (*e.g.*, direction, amplitude). Another important difference is that, until now, there were no extensions and applications of the STATIS method to the study of social networks. The fourth most similar work to ours is the one by Skillicorn et al. (2014). These authors proposed a spectral approach for modelling and analysing directed networks that change over time. Their idea is similar to ours in what regards modelling of a dynamic social network and the procedure to analyse them. The evolving social network is modelled by first bonding together a series of network snapshots into a single data structure, and then applying an eigendecomposition in order to produce a common space where the evolution of nodes and their relationships can be

analysed. Like us, they also generate trajectories to examine the evolution of nodes. However, there are important differences between this and our work. First, while we model the dynamic network using three-order tensors embedded with structural information, they first aggregate all adjacency matrices into a single adjacency matrix with temporal information and then convert it into a random walk Laplacian matrix, whose meaning is harder to grasp. Then, a specific two-way eigendecomposition developed for undirected graphs is applied so as to generate the common space where the nodes and their trajectories are projected. In turn, we resort to three-way methods instead of two-way matrix decompositions, and directly embed the temporal and the relational information of the dynamic network into one unique data structure (the three-order tensor), which makes our methodology more parsimonious. Besides, the common space generated using our procedure can be interpreted according to the measures used to describe the entities, potentially giving more useful information to the analyst. Somewhat related to our work is the methodology proposed by Sarkar and Moore (2005). These authors introduced a two-stage methodology for modelling dynamic social networks based on latent space models. In the first stage, the authors use a time variant extension of the classical MDS, developed by them, in order to obtain initial estimates of the entities' positions in a p-dimensional Euclidean latent space (in their experiments $p = 2$). The distance matrix required by MDS is generated by computing the geodesic distance among entities (*e.g.*, nodes) in the network. After this batch initialisation of the entities' positions for all time steps, these coordinates are updated using a forward method in the second stage. The goal of the forward method is to learn the new entities' coordinates at each time step in a way that ensures that these coordinates reflect the current network and, at the same time, do not deviate strongly from the coordinates obtained in previous time steps. Although they also represent the coordinates associated with each time step of a given node in a bidimensional space, one important difference is that the information displayed only gives clues about changes in the proximity among nodes with respect to geodesic distances. By embedding structural information about the network dynamics in our data model, our methodology is able to provide a richer overall picture of the changes taking place at different levels of the network.

4.4 Methodology

In this section we discuss two alternative matrix-based models of social networks and argue which one is more appropriate for representing dynamic networks. We also introduce our visualisation methodology, which is based on the generation of spatio-temporal trajectories of nodes, or communities, in low-dimensional spaces. These trajectories are obtained by exploring the compressed information produced by an estimated Tucker3 model.

4.4.1 Modelling Dynamic Social Networks as Three-order Tensors

As motivated in the previous sections, we aim at producing a meaningful and compact visual representation of the evolution of nodes, or communities, in a dynamic social network comprised of a finite number of snapshots. By *meaningful* and *compact* we mean that this representation is not only able to be grasped by human eye, through projections in low-dimensional spaces, but it is also focused on the relevant patterns contained in the structure of the dynamic network. This is an important aspect that guided our understanding of how a social network should be modelled in order to meet requirements such as interpretability of results. In fact, if we choose to model a social network, comprised of n nodes, by means of an adjacency matrix $A_{n\times n}$ (*e.g.*, this strategy was employed by Bader et al. (2006), Dunlavy et al. (2011), and Bach et al. (2014)), and factorize it, we obtain new entities that are difficult to interpret, especially if we are dealing with directed networks, where the relationships among entities are asymmetric. In such cases, the rows are related to the out-neighbourhood and the columns are related to the in-neighbourhood of nodes. On the other hand, when working with symmetric adjacency matrices, the new entities returned by the decomposition are typically the same for both modes A and B, thus creating redundant results and less valuable information than one would expect from rich network structures. To overcome the barriers posed by the adoption of a standard social network representation model, we suggest computing SNA measures (*e.g.*, degree centrality, eigenvector centrality, closeness centrality, betweenness centrality, and clustering coefficient), to embed richer structural information into the *network snapshots* (*i.e.*, matrices that represent snapshots of the state of the network for a specific time step). In the context of social networks, these measures are useful in the sense that they give a high-level description of the network structure, thus offering insight about the position and importance of each entity in the network. Moreover, this strategy helps mitigate the differences between directed and undirected networks and improves the interpretability of the component matrices yielded by the Tucker3 model.

Thus, we model a dynamic social network as a series of snapshots, each providing the state of the network at a particular moment in time. For each one of these network snapshots we compute a set of SNA measures and store this information as a *snapshot matrix*, where rows are associated with nodes (mode A) and the columns with the computed measures (mode B). Note that some nodes may only be active at specific moments of the network evolution. To handle these cases, we assume that they are always present, *i.e.*, there is a row in all snapshot matrices populated with information about these nodes. We consider this to be a weak assumption since the absence of a node at a given time step is reflected in zero matrix entries for the SNA measures. These zero entries indirectly capture the absence of these nodes from the network by flagging that these nodes are not connected to other nodes during these time steps. After using this strategy to build the snapshot matrices, which are embedded with structural information regarding each snapshot of the social network, the process of converting them into a three-order tensor becomes trivial. To do so, one just needs

to introduce the additional mode C, associated with the temporal dimension, and bind together the temporally aligned snapshot matrices into a single data structure, thus obtaining a three-order tensor $\chi \in \mathbb{R}^{I \times J \times K}$, where I ($i = 1, ..., I$) denotes the number of entities of the first mode A (the row-entities, which are defined along the horizontal axis), J ($j = 1, ..., J$) refers to the number of entities of the second mode B (the column-entities, which are defined along the vertical axis), and K ($k = 1, ..., K$) indicates the number of entities of the third mode C (the fiber-entities, which are defined along the depth axis). In the social networks context, the mode A is the dimension of nodes, mode B is the dimension of node-level SNA measures, and mode C is the temporal dimension.

4.4.2 Spatio-temporal Trajectories

A spatio-temporal trajectory can be defined as a sequence of K time-stamped points $TR = p_1 \rightarrow p_2 \rightarrow, ..., \rightarrow p_i \rightarrow, ..., \rightarrow p_K$, where $p_i = (x_i, y_i, t_i)$ ($i = 1, 2, .., K$) represents the coordinates of a given entity in a 2D space, at time step t_i. It is common to choose 2D, instead of 3D representations, because they are simpler to understand and analyse and, at the same time, allow for an effective data analysis. Thus, we use two-dimensional projections and encode the temporal dimension as a trajectory over the plane. This way, we are able to map a given node's, or community's, movement across time in a static bidimensional space.

To better understand the concept of spatio-temporal trajectory consider an entity, which can be a node or a community, that evolves over five time steps in a bidimensional space spanned by variables x and y. A spatio-temporal trajectory graphically displays the evolution of this entity in this specific bidimensional space, as can be ascertained from Figure 4.2.

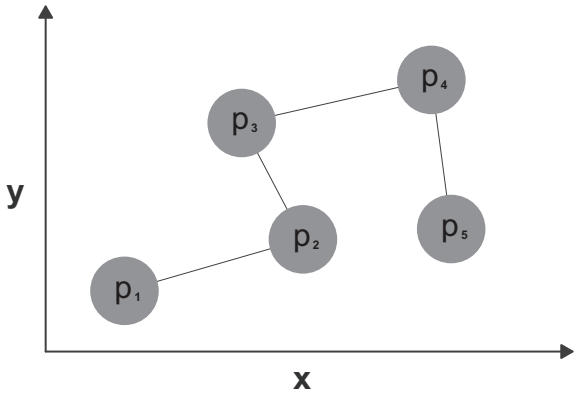


Figure 4.2: Illustration of the spatio-temporal trajectory of an entity over five time steps.

An important requirement for generating meaningful trajectories of entities is the existence of a single common space for all time steps K where the trajectories can be projected (Skillicorn et al., 2014). Such a space enables a rigorous analysis of the state of an entity at consecutive time steps, as well as the comparison of trajectories of distinct entities. We ensure that this requirement is met

by first binding the several states of the dynamic social network into a single data structure (*i.e.*, the three-order tensor), and then decomposing this single structure. The result of the decomposition is a set of component matrices, each associated with a spectral space defined on each mode of the tensor. Since the decomposition is performed simultaneously over all snapshots of the dynamic social network, and for all entities, each one of the resulting component matrices defines a common space where the evolution of these entities can be projected and analysed.

Node-Level Trajectories

In order to generate the spatio-temporal trajectories of each node we decompose the three-order tensor, comprised of a set of temporally ordered snapshot matrices, by estimating a Tucker3 model. Then, we consider the bidimensional space spanned by the two most representative components of matrix **B**, as given by the entries of the core tensor $\mathcal{G}$, and define the x and y coordinates for each time step k ($k = 1, ..., K$) of the trajectory. We obtain these coordinates for each node i ($i = 1, ..., I$), by computing the dot product between $x_{i,:,k}$ (horizontal fibers of χ) and each column of matrix **B** associated with the two components that explain the highest portion of the total data variance. These two components are assigned to the x-axis and y-axis, respectively. This vector operation returns the coordinates, or the bidimensional position vector, for each node i of the network. This process is depicted in Figure 4.3.

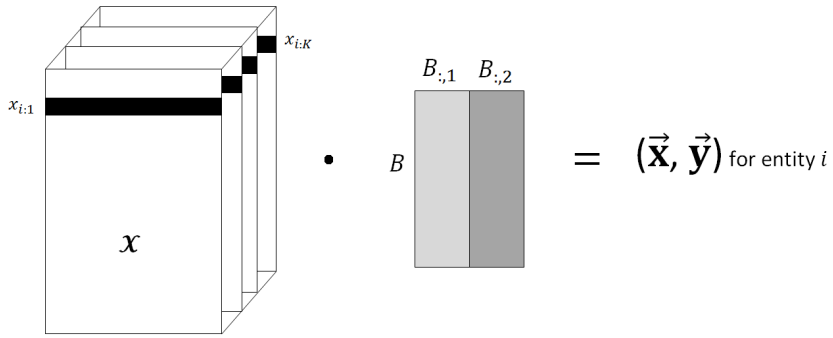


Figure 4.3: Illustration of the procedure to obtain the coordinates of the spatio-temporal trajectory of a given entity i pertaining to mode A. Being a trajectory defined as a set of pairs (x,y) linked according to the given temporal order, we obtain each one of these coordinates by computing the dot product between each row vector of χ, associated with entity i, and the two columns of component matrix **B** explaining the greatest variation in data. In this example, the trajectory of entity i will be defined by a set of K ($K = 4$) coordinates (x,y), each one corresponding to a given time step k.

In the last step the evolution of each node is analysed in order to obtain a qualitative description of an entity's trajectory. The movement, or trajectory, of a given node can be characterised by a direction (*e.g.*, upwards, downwards, leftwards, rightwards), which can be more regular or more irregular; and by an amplitude, which can be higher, thus covering a larger space area, or lower, by keeping its position in the plane almost unchanged over time. We will take these features into account when analysing the trajectories in the case study (Section 4.5).

In short, our visualisation approach maps each snapshot of the social network of a given entity (node or community) into a point in a common Tucker space, and links these points in order to define a spatio-temporal trajectory that represents the dynamic behaviour of this entity across time. Figure 4.4 outlines the steps of our methodology for generating trajectories of evolving nodes.

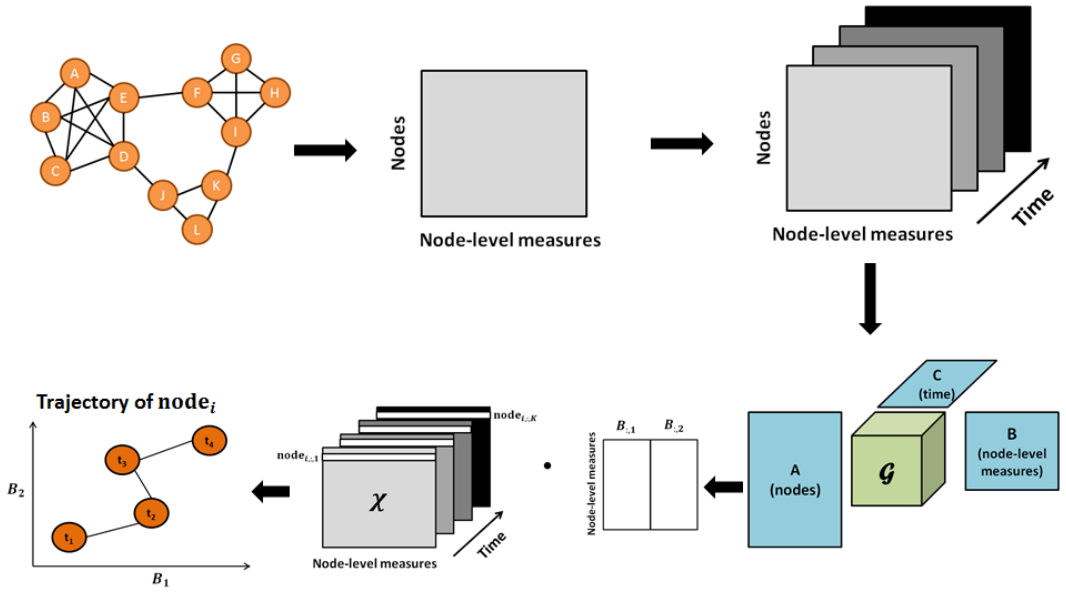


Figure 4.4: Overall view of the steps of the methodology for generating node-level trajectories. The first step is to convert each snapshot of a dynamic network into a snapshot matrix. Then, these matrices are arranged into a three-order tensor in a way that preserves their temporal order. The tensor is further decomposed using a Tucker3 model with orthogonality constraints. The dot product between the two most representative components of matrix **B** (denoted as $B_{:,1}$ and $B_{:,2}$ in the figure) and the row vectors associated with a given entity i, gives the (B_1, B_2) coordinates of node i in the common 2D space of matrix **B**. These coordinates define the spatio-temporal trajectory of node i. The axes of the space spanned by the two components of **B** are labelled according to their meaning in order to allow for the semantic interpretation of the movement of the trajectory.

Although in this chapter we restrict our analysis to a very small network, in theory our approach also holds when dealing with large social networks. Sun et al. (2008) proposed an incremental algorithm to compute the Tucker decomposition for large tensors. To estimate the adequate number of components of the Tucker3 model, instead of using heuristic methods, such as the DIFFerence in FIT (DIFFIT) (Timmerman and Kiers, 2000), one can use more efficient methods, such as the Automatic Relevance Determination (ARD) framework (Mørup and Hansen, 2009).

Community-Level Trajectories

The previous procedure can be extended to visualise the dynamics of mesoscopic structures, such as communities of nodes. Studying the evolution of communities is especially relevant when dealing with networks comprised of hundreds, thousands, or millions of nodes. In these cases, studying the evolution of all nodes in a network using our methodology is impractical and prohibitive. So,

unless one is interested in analysing one specific node, or a bounded set of nodes, increasing the level of abstraction to communities can ease the analysis, while providing a higher-level view of the structural changes occurring in the network. Notwithstanding, both types of trajectories can be used as complements in the analysis of evolving social networks, rather than substitutes.

The distribution of ties in social networks is often inhomogeneous (Fortunato, 2010). As a consequence, social networks typically display community structure (Newman, 2003b). These communities are commonly defined as groups of highly interconnected nodes, which are sparsely connected among them. Examples of such communities include circles of friends, families or work teams. The identification and analysis of these groups is important as it offers an effective way to summarise the broad patterns in the network structure, providing insights into the organisation of the underlying social system (Aggarwal and Subbian, 2014). Since these are substructures of the network, the evolution of communities reflects the changes taking place at the network (*e.g.*, in a social network this may happen due to, for instance, changes in the role, social activity, and communication patterns of actors). Hence, communities are patterns that tend to be highly unstable over time, in the sense that they can disappear, split into smaller communities, merge into larger communities, contract, grow and so on. The temporal sequence of such critical evolutionary events characterises the communities' life-cycle and has implications in the way trajectories are displayed in a latent 2D space. Thus, our visualisation approach has to be adjusted so as to provide the viewer with a faithful portrait of the evolution of the community, which depicts these critical evolutionary events and gives clues about the reasons that may explain them.

Given this, the first challenge is how to represent a dynamic community that may take different paths during its evolution as a consequence of undergoing splits or merges during its lifespan. In this work, we resort to the MECnet framework (Oliveira et al., 2014), which was devised to monitor dynamic communities over time, to address this challenge. MECnet models dynamic communities using an evolution graph, defines them using the temporal sequence of community instances that comprise them, and summarises their evolution by describing their life-cycle. The temporal sequence of community instances associated with a given dynamic community is termed *temporal trajectory* and encapsulates the different trajectory paths taken by the community during its evolution. As implied by the distinct designations, there is a difference between MECnet temporal trajectories and the spatio-temporal trajectories defined in this work. In the former, trajectories are defined as a sequence of labels indicating the community instances' ID and the corresponding time steps, whereas in the latter trajectories are associated with (x, y) coordinates indicating the position of the community instances in a 2D Tucker space. Thus, the definition of trajectory adopted in this work enriches MECnet trajectories by associating them to a spatial context. This spatial context allows to express the changes detected by MECnet in a more comprehensible way by assigning a meaning to them. In Figure 4.5, we show an example of the MECnet evolution graph depicting the evolution of three dynamic communities (Com^A, Com^B,

and Com^C) over a time span of four time steps. Taking dynamic community Com^C as an example, its life-cycle is given by $LC_{Com^C} = \{born_{[t_1]}, survival_{[t_1,t_2]}, split_{[t_2,t_3]}, survival_{[t_3,t_4]}\}$ and the temporal trajectory of this community is defined by the following sequence of community instances $TT_C = \{Com^3_{t_1}, Com^3_{t_2}, Com^3_{t_3}, Com^3_{t_4}, Com^4_{t_3}, Com^4_{t_4}\}$.

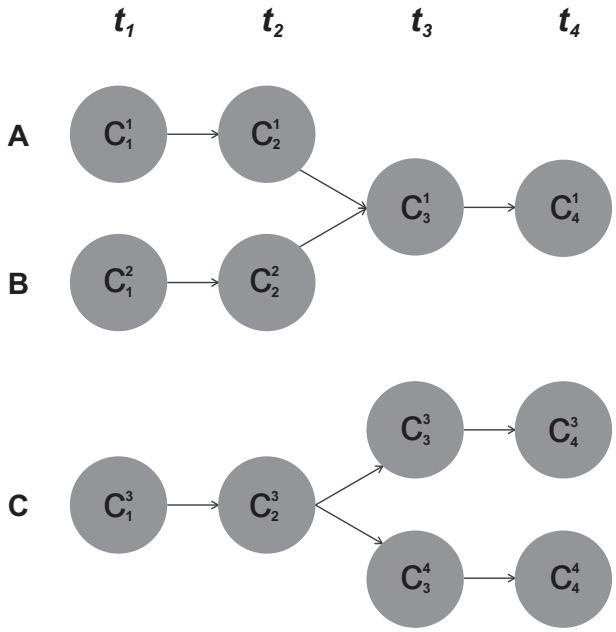


Figure 4.5: Evolution graph depicting three dynamic communities tracked over four time steps, featuring merging and splitting life-cycle events.

As illustrated in the previous example, a dynamic community may take two or more trajectory paths, as a consequence of split events. Since a dynamic community is assumed to be a single entity, these alternative paths pose problems during the modelling step. In order to ensure data consistency, our data model requires that each row (or column or fiber) of the three-order tensor represents the same entity, regardless of the time step considered. However, the representation of a dynamic community with several paths would require as many rows as the number of paths. For instance, to model the dynamic community Com^C depicted in Figure 4.5 two rows would be needed to accommodate the two paths generated by the split event. To address this challenge, we employ the same strategy used for modelling evolving nodes, by arranging a set of temporally aligned snapshot matrices into a three-order tensor. The rows of the tensor are associated with the lowest-level entities in the network, *i.e.*, the nodes, in order to avoid the modelling problem of dynamic communities. The columns of the tensor represent the SNA node-level measures and the fibers represent time. A Tucker3 model with orthogonality constraints is then applied to this tensor. Until now, the methodology steps are exactly the same as in the node-level case. The adaptations to the community-level case are introduced in the following steps. Besides the outputs of the tensor decomposition, additional information is required to proceed with the methodology.

One of the most important steps is the discovery of meaningful communities at each snapshot of the network using a community detection algorithm. Due to its desirable features, in this study we resort to the Louvain method (Blondel et al., 2008) to identify the communities of nodes and induce a higher-order structure of the network. The main advantages of this method are being parameter-free, very fast, and able to detect high-quality network partitions. One of its drawbacks is being an order-sensitive algorithm, returning different results according to the order of the nodes in the sequential analysis of modularity gains. Another shortcoming is being sensitive to noise. After identifying the communities in each network snapshot, MECnet is applied to these results in order to obtain the temporal trajectories associated to each dynamic community, as well as the community membership of each one of the community instances that comprise these trajectories. Having both these information and the results of the Tucker decomposition, an aggregation step is carried out aiming to extract the coordinates of the spatio-temporal trajectories of each dynamic community. The aggregation is performed as follows: for each dynamic community, (i) aggregate the values of each SNA node-level measure for the set of nodes belonging to each community instance, using the information provided by the original three-order tensor, the MECnet temporal trajectory and the community membership vector; (ii) extract the (x,y) coordinates of each community instance by computing the dot product between the row vector containing the aggregated measures for the community instance and the column vectors associated with the two most representative components of matrix **B**; (iii) repeat procedure (ii) for each time step and, if necessary, for each community instance, since we may have two or more community instances associated with a given time step. For the sake of simplicity, in this work we use the mean as the aggregation measure, although other measures can be used. After conducting the described procedure, the spatio-temporal trajectory of each dynamic community is obtained by projecting the coordinates obtained for the corresponding community instances in the 2D space spanned by the components of matrix **B**. Then, the interpretation of the community trajectories is performed by following the same guidelines presented for the analysis of node trajectories (*i.e.*, by analysing the direction and amplitude of the trajectory). We provide an illustrative example of a community-level trajectory in Figure 4.6. The life-cycle of this community is $LC_{Com} = \left\{born_{[t_1]}, survival_{[t_1,t_2]}, split_{[t_2,t_3]}, survival_{[t_3,t_4]}\right\}$ and the dynamic community is defined as $TT_{Com} = \left\{Com^4_{t_1}, Com^3_{t_2}, Com^3_{t_3}, Com^1_{t_4}, Com^1_{t_3}, Com^3_{t_4}\right\}$. Figure 4.7 summarises the steps of the introduced methodology.

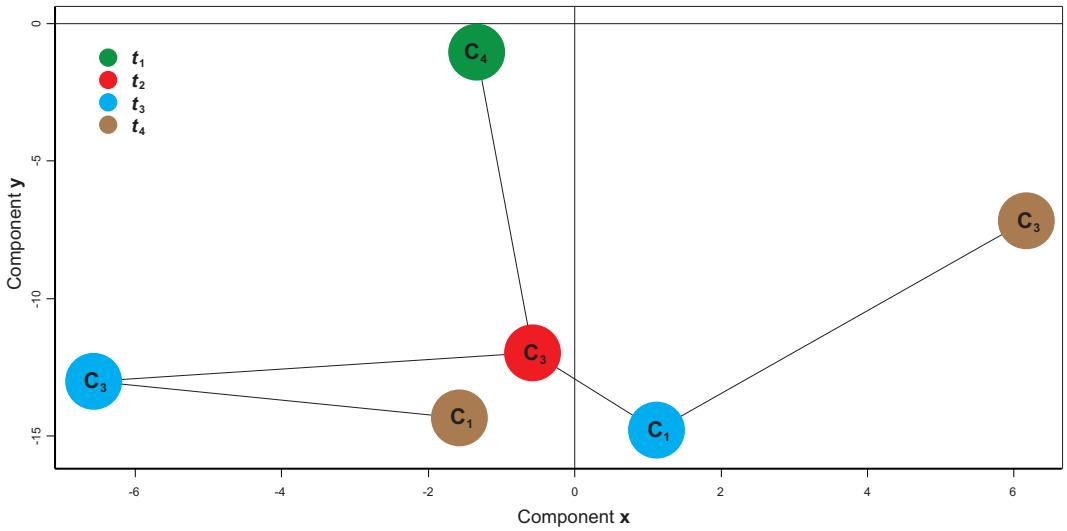


Figure 4.6: Spatio-temporal trajectory of a dynamic community over four time steps $[t_1, t_4]$, in the space spanned by the two most representative components of matrix **B**.

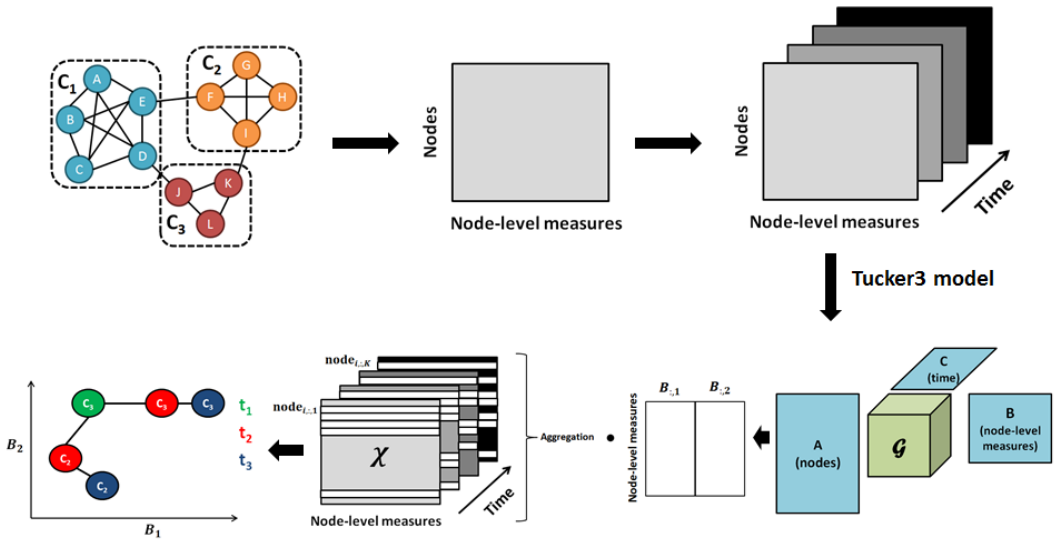


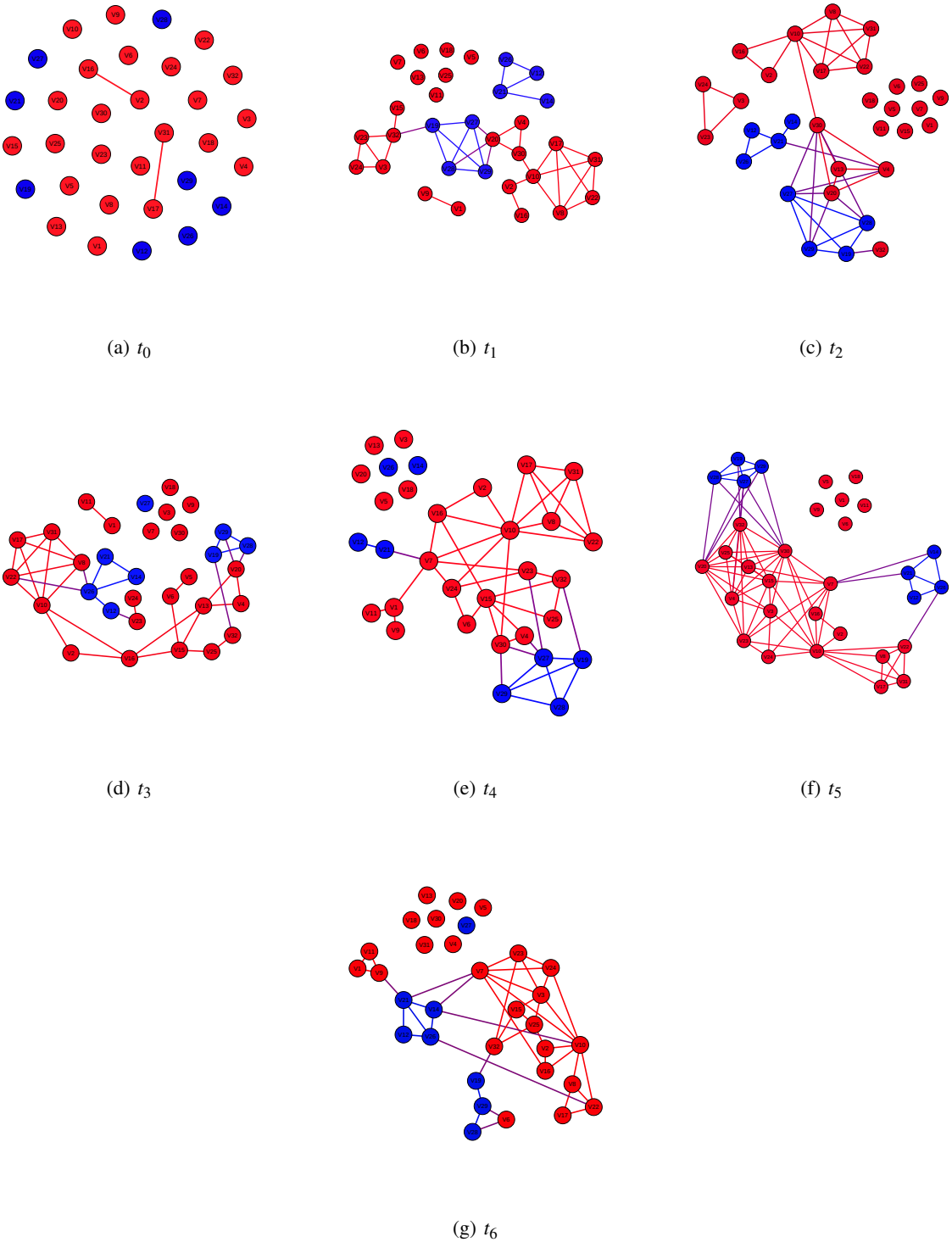
Figure 4.7: Overall view of the steps of the methodology for generating community-level trajectories. The modelling procedure is the same introduced for node-level trajectories. The main differences are: (i) an additional aggregation step and (ii) the output. The former takes the information provided by MECnet to identify the set of community instances associated with each dynamic community and the set of nodes pertaining to each community instance. The aggregation of the SNA node-level measures values is then performed over these sets of nodes. The latter uses the MECnet temporal trajectory of each dynamic community to project the coordinates of the corresponding community instances in the Tucker space, as well as their paths. These paths provide a visual representation of the sequence of critical evolutionary events undergone by the community over time.

4.5 Case Study

This section presents an illustrative case study using a set of friendship networks collected at seven different moments in time, for the same individuals. The goals of this case study are: (i) to verify the suitability of the proposed visualisation approach for representing the dynamics of social networks and (ii) illustrate the step-by-step procedure of our methodology and explain how it can be used in practice. The methodology is comprised of the following steps: first, we select and compute the SNA node-level measures that describe the social entities; secondly, we model the resulting snapshot matrices as a three-order tensor, where the third dimension is time; the third and fourth steps consist in preprocessing the data and applying the Tucker3 model to the three-order tensor, respectively; finally, the 2D coordinates of each entity's trajectory are computed and interpreted based on the results of the model. We go through each one of these steps in the subsequent subsections.

4.5.1 Data Description

The dataset used in this case study comprises several snapshots of a temporal friendship network. These network data was collected by Van De Bunt et al. (1999) among a group of university freshmen enrolled in a common study program in a Dutch university. The aim of the data collection was to investigate the important factors behind friendship formation. Data were gathered by questionnaires which were delivered to 49 students in seven different occasions asking them to indicate and rate their social relationships using a six-point scale. These occasions are not equally spaced in time, since the first four time steps are three weeks apart, and the last three time steps are six weeks apart. There were also changes in the number of students, which drops from 49 to 32, due to university drop-outs and non-response of questionnaires in at least four of the total seven occasions. The final group was comprised of 24 female and 8 male students. It is also important to note that, with a few exceptions, the respondents were mutual strangers (freshmen) when the study first began, which reflects in a sparse adjacency matrix for the first measurement (t_0). In the original dataset the ties linking each pair of students were coded according to a specific scale, which represented the rating of the established relationships. These ties were also directed, since the data were collected by questionnaires. The reciprocity, *i.e.*, the fraction of mutual connections in the directed network, for all time steps (occasions) was on average $r = 0.6$ ($\sigma = 0.07$). For the sake of simplicity, we converted the original network into an undirected binary network. To do so, we kept only the mutual ties and removed the asymmetric ones. We also replaced codes $6 =$ *item non-response*, $9 =$ *actor non-response*, $4 =$ *neutral*, and $5 =$ *troubled relation* by 0 and codes $1 =$ *best friend*, $2 =$ *friend*, and $3 =$ *friendly relation* by 1. Therefore, the absence of a tie, coded by 0, represents one of the following situations: (i) absence of relationship between a given pair of students, (ii) non-response, or (iii) the lack of intention to establish a friendship on the part of at least



(a) t_0 (b) t_1 (c) t_2

(d) t_3 (e) t_4 (f) t_5

(g) t_6

Figure 4.8: Node-link representation of the undirected and unweighted Van de Bunt's temporal friendship network at seven different occasions. Each graph maps the network of friendship among a set of freshman students at a given moment in time. Nodes are coloured according to students' gender: red nodes represent female students and blue nodes represent male students.

one of the students. The existence of a tie, coded by 1, means that there is a friendly relationship between the corresponding pair of students, which reflects in a stronger willingness to interact with each other. Formally, each entry of the adjacency matrix $A_{n\times n}$, associated with time step t_k ($t = 1, ..., K$), can take the following values: $a_{ij} \in \{0, 1\}$, thus being a binary matrix. The number of non-responses for each student belonging to the final set is given in Table 4.1. The node-link representation of the undirected and unweighted version of these friendship networks is depicted in Figure 4.8. Besides sociometric data, additional information is available regarding the students' gender (male or female), smoking behaviour (smoker or non-smoker), and education program (regular 4-year program, special 3-year program or special 2-year program). These attributes are useful for the further interpretation and understanding of the results.

Table 4.1: Number of non-responses for each student of the final set (32 students) and the corresponding time points.

Students	Non-Responses	Time points	Students	Non-Responses	Time points
S1	2	t_2, t_5	S17	0	–
S2	0	–	S18	2	t_1, t_4
S3	2	t_3, t_4	S19	0	–
S4	1	t_6	S20	2	t_4, t_6
S5	2	t_5, t_6	S21	0	–
S6	0	–	S22	0	–
S7	1	t_3	S23	0	–
S8	0	–	S24	0	–
S9	1	t_3	S25	1	t_2
S10	0	–	S26	1	t_4
S11	2	t_1, t_2	S27	2	t_3, t_6
S12	0	–	S28	0	–
S13	2	t_4, t_6	S29	0	–
S14	1	t_4	S30	2	t_3, t_6
S15	0	–	S31	1	t_6
S16	0	–	S32	0	–

The reasons behind the choice of these network data are: (i) temporal dimension, which makes it possible to track changes over time in students' social behaviour, (ii) stable set of entities, since the same set of students is monitored over time, (iii) small social network, thus suiting better the illustrative purpose of this case study, and (iv) availability of additional information concerning students' attributes, which helps to validate the results in the absence of a ground truth.

In order to describe and understand the general configuration of the temporal friendship network, and derive some information about the evolution of its structure, we compute classical network-level measures, such as density, diameter, number of connected components, average degree centrality, average geodesic distance, global clustering coefficient, and size of the largest clique, for each network snapshot. The obtained values are reported in Table 4.2. The network under analysis

Table 4.2: Descriptive statistics and values of the network-level measures obtained for the Van de Bunt's temporal friendship network, at seven different occasions.

Network-Level Measures	t_0	t_1	t_2	t_3	t_4	t_5	t_6
Number of Nodes	32	32	32	32	32	32	32
Number of Ties	2	36	42	38	46	81	43
Number of Connected Components	30	10	11	8	8	7	9
Size of the Largest Component	2	19	20	24	25	26	24
Global Clustering Coefficient	-	0.61	0.65	0.52	0.4	0.57	0.4
Size of the Largest Clique	2	4	5	5	5	7	4
Average Degree Centrality	0.12	2.25	2.62	2.38	2.88	5.06	2.69
Density	0	0.07	0.08	0.08	0.09	0.16	0.09
Average Geodesic Distance	-	3.65	2.78	3.86	2.71	2.19	3.3
Largest Diameter of the Connected Components	-	9	5	10	5	5	8

captures the relationships among 32 students and these relationships are monitored over seven time steps. At the beginning of the academic year, the majority of freshmen were not acquainted with each other, as indicated by the extremely low number of ties at time step t_0, by the very large number of connected components (the majority corresponds to singletons), and by the low values of density and average degree centrality. This scenario changes after three weeks since the number of friendship ties sharply increases from time step t_0 ($m_{t_0} = 2$) to time step t_1 ($m_{t_1} = 36$), suggesting that $[t_0, t_1]$ was a time period of intense friendship formation. The number of ties continues to rise, although at slower rates, until t_5. Then, from t_5 to t_6 the number of ties drops from 81 to 43. From the global analysis of the table, we conclude that there is a stark contrast between time steps t_0 and t_5 as these represent the most extreme cases in the evolution of the friendship network. While in t_0 almost no student knows each other, which reflects in a scattered network, in t_5 the network connectivity and reachability reaches its peak, as deduced from the values of density, average degree centrality, and average geodesic distance, at t_5. The remaining time steps are more similar between them. The analysis of the network-level measures also indicates the presence of a large connected component that comprises the majority of students in all time steps, other than t_0. Inside this large component, students form tightly-knit friendship groups where every member is connected to every other by friendship ties. The global clustering coefficient suggests that the presence of such *cliques* is very frequent in the network, reaching its peak in t_2. It is also interesting to observe that, as expected, the size of the largest clique is not very high, being on average comprised of five students. In other words, students tend to develop close bonds with only a few number of people, due to the investment required by close friendships. This idea is corroborated by the low values of the average degree centrality. The analysis of this measure reveals that the average number of friendship ties oscillates between two and five and does not change much over time. When these relationships are established outside the local group of friends, the average geodesic distance among students and the

network diameter decline as it becomes easier to reach other people in the network through a fewer number of links. These two last measures indicate, for instance, that in t_2 and t_4 approximately three connections are needed to reach anyone in the network and, in the worst case, this number of connections increases to five. These low values of the average geodesic distance and the diameter reveal the social proximity among the freshmen students, which is fostered by the sharing of the same social context.

4.5.2 Measures Selection, Preliminary Tests, Preprocessing and Tucker3 Model Generation

In order to obtain the spatio-temporal trajectories for each freshman student, we first need to transform the initial adjacency matrices into snapshot matrices embedded with structural information about the individuals. To do so, first we need to decide which SNA measures should be used to characterise these individuals. In this case study, we select the normalised versions of the following node-level measures: degree centrality, eigenvector centrality, closeness centrality, betweenness centrality, and clustering coefficient, because they capture different dimensions of the centrality concept. Notice, however, that the proposed methodology is not constrained by this specific set of measures and can incorporate other types of variables, contingent on both the goal of the analysis and the nature of the networks at hand. Here we provide a brief description of the selected measures within the context of the students' friendship networks:

- **Degree Centrality**: number of mutual friendly relationships of a given student;
- **Eigenvector Centrality**: ascribes a relative score to each student that measures how well he/she is connected to other well connected students in the network. A student with high degree centrality and high eigenvector centrality is assumed to be popular in the network, because he/she has many friends and, at the same time, some of these friends are also popular in the network;
- **Closeness Centrality**: measure of reachability that measures how fast can a given student reach everyone in the network;
- **Betweenness Centrality**: measures the extent to which a given student lies between other students in the network. Students with high betweenness centrality occupy critical roles in the network structure, since they usually have a network position that allow them to work as a bridge between tightly-knit groups, being vital elements in the connection of different regions of the network;
- **Clustering coefficient**: the local version of this measure quantifies the transitivity in the neighbourhood of a given student, by indicating the level of cohesion between his/her neighbours.

Although the proposed visualisation focuses on observing trajectories of single nodes (or single communities), some of the measures we use to describe these nodes depend on the global network structure (*e.g.*, eigenvector centrality, closeness and betweenness centralities). The use of such measures allow us to capture latent properties of the whole network, in addition to structural information about the nodes themselves, when reducing the original measures' space into a bidimensional one. As a consequence, changes in the nodes' trajectories also reflect changes in the global network structure, and not only changes at the node-level. Such global changes can be related to, for instance, the increase/decrease of the number of connections of a given node in the network that is not directly connected to the node under analysis.

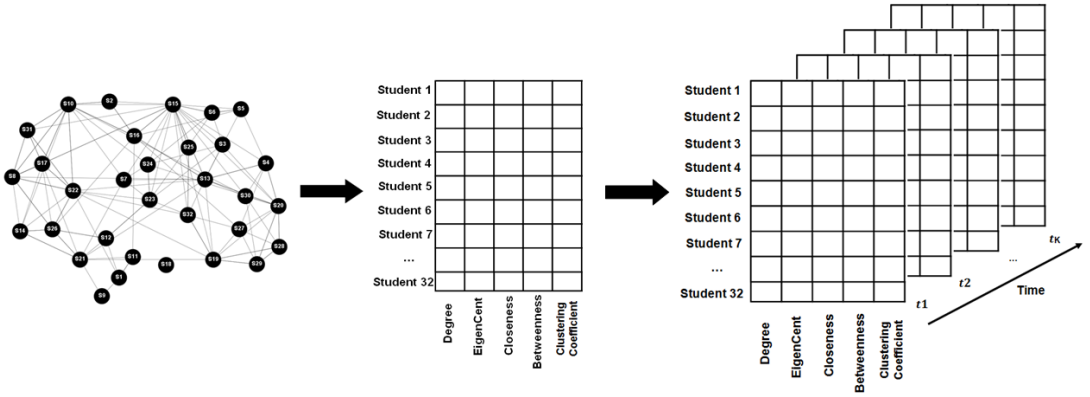


Figure 4.9: Illustration of the process of converting a social network that evolves over time into a three-order tensor. Given a snapshot of an evolving social network, SNA measures are computed for each node in V and the resulting data are organised as a snapshot matrix. These steps are repeated for all available snapshots of a network. The set of matrices produced by these steps is bonded together into a three-order tensor.

After selecting the measures we organise the snapshot matrices as a three-order tensor $\chi \in \mathbb{R}^{32\times5\times7}$, where the first mode encompasses the 32 students, the second mode refers to the five SNA measures, and the third mode is related to the seven different occasions or time moments when the questionnaires were applied. Each entry of this tensor gives the score of student i ($i = 1,...,I$) on node-level measure j ($j = 1,...,J$) at measurement occasion k ($k = 1,...,K$). The process of converting the set of network snapshots into a three-order tensor is represented in Figure 4.9.

Since three-way analysis is only recommended in cases where the data contain a non-negligible three-way interaction across the three data modes, or at least two non-negligible two-way interactions, we carry out a simple fixed-effects three-way ANalysis Of VAriance (ANOVA) on the resulting three-way array $\chi \in \mathbb{R}^{32\times5\times7}$, using the three different modes as factors. The goal is to assess the effect size of all variance components, where the total variance to be explained is the variance of all data elements with respect to the grand mean (Kiers and Van Mechelen, 2001). This step is important to understand if a two-way analysis (*e.g.*, PCA, SVD) on data averaged over one of the modes (*e.g.*, the students mode) would suffice to capture the main patterns contained in data. Only the cases where a substantial effect size of a three-way interaction, or two or more two-way

interactions, are present warrant the use of a three-way analysis (*e.g.*, the Tucker3 model). The results of the fixed-effects three-way ANOVA for the friendship networks are reported in Table 4.3. From the analysis of the decomposition into sum of squares of the main effects, two-way interactions, and three-way interaction (plus error term) given by this table, we deduce that the highest contribution derives from a sizeable three-way interaction (plus error term) established between all three modes. The second highest contribution is associated with the two-way interaction established between the students mode and the time dimension.

These results testify to the presence of a substantial three-way interaction (plus error term) that can only be captured by means of a three-way analysis, thus validating the use of a Tucker3 model. Moreover, the large effect size of the node-level measures term suggests the appropriateness of using mode B, rather than other modes, to generate the space where the trajectories will be mapped.

Table 4.3: Three-way ANOVA of Van de Bunt's network data, after subtraction of the grand mean, with *students*, *node-level measures*, and *occasions* (or time) as fixed factors. The highest contributions are highlighted in bold.

Effects	**Sum of Squares**	**Percentage**
Students	11.79	9.12%
Node-level measures	18.5	*14.27%*
Occasions	4.63	3.58%
Students × Node-level measures	14.53	11.24%
Students × Occasions	20.49	**15.84%**
Node-level measures × Occasions	9.05	6.99%
Students × Node-level measures × Occasions + Error	50.4	**38.96%**

After checking if the data at hand warrant the conduction of a three-way analysis, we move to the second step of the methodology: data preprocessing. We preprocess the three-way array $\chi \in \mathbb{R}^{32\times5\times7}$ by first centring it across mode B, and then scaling it within mode B. While centring is a projection step that removes the offset terms that represent the unknown neutral points, scaling adjusts scale differences and accommodates heteroscedasticity. To perform these steps, we follow the guidelines of Bro and Smilde (2003). Thus, centring was done by first rearranging the array into a two-way matrix $J \times IK$, and then centring this matrix across mode B as in ordinary two-way analysis. Scaling was carried out by dividing the values of each variable (entities of mode B) by the square root of the sum of squares of the values of all students in all occasions, on this variable. The scaling process equalises the importance of the variables (*i.e.*, the node-level measures) in the analysis, by eliminating the effect of scale range differences.

In the next step of our methodology we decompose the three-order tensor into a small core tensor $\mathscr{G}$ and a set of component matrices **A**, **B**, and **C**, that summarise the entities in the three modes. To choose among the myriad of possible solutions for the Tucker3 model, we created a model for each combination of components $(P \times Q \times R)$, where P, Q, and R are bounded above

by five components (*i.e.*, the maximum possible order tested was $(5 \times 5 \times 5)$), which is the lowest number of entities in all three tensor modes. The fit values associated with each model were then computed for each possible order. We selected the order of the model using the DIFFIT method proposed by Timmerman and Kiers (2000), which is a variant of the scree test for three-way analyses. This method finds the number of components P, Q, and R ($P < I, Q < J$, and $R < K$), by balancing model fit (*i.e.*, amount of information accounted for) and model parsimony (*i.e.*, low P, Q, and R). The steps involved in the DIFFIT method are the following: (i) run the Tucker3 model for all possible combinations of components up to a maximum possible order (in our case, $(5 \times 5 \times 5)$); (ii) compare the fit of the solutions comprising the same total number of components $P+Q+R$ (also known as *model complexity*), and keep the solution with the largest fit; (iii) after carrying out the previous step for all possible $P+Q+R$, list the corresponding best solutions (*i.e*, those with largest fit), sort them in descending order of $P+Q+R$, and compute their consecutive differences in fit; (iv) select the model for which this difference in fit is maximum. Based on this procedure, we selected and further estimated an orthogonality-constrained Tucker3 model of order $(4 \times 3 \times 4)$, which explains 82.65% of the total data variation. This order is a parameter of the model and refers to the number of components retained in each mode ($P = 4$, $Q = 3$, and $R = 4$). We have used the `ThreeWay` package (Giordani et al., 2012) for `R` environment (Team, 2008) to perform the Tucker3 tensor decomposition and the `igraph` (Csardi and Nepusz, 2006) package to compute the node-level measures. The full methodology, which includes the generation of spatio-temporal trajectories, was implemented in the `R` language.

4.5.3 Interpretation of the Coordinate Axes of a Tucker3 Space

After decomposing the three-order tensor, we analyse the output of the resulting Tucker3 model. One important step within the scope of our methodology is the interpretation of the meaning of the coordinate axes of the maximum-variance 2D common space where the trajectories will be projected. The procedure we use for the interpretation of the 2D space is independent of the level of analysis (micro-level or meso-level). The interpretation is ideally done by observing each component as a latent variable and then assigning a label that expresses its meaning. However, and similarly to two-way analyses (*e.g.*, PCA, SVD), the interpretation of these components as theoretical constructs is not straightforward and is highly prone to subjectivity (Kiers and Van Mechelen, 2001; Kroonenberg, 2008). To mitigate this problem, it is common practice to rotate the component matrices using procedures such as the varimax orthogonal rotation. The rotational freedom of the Tucker3 model allows the exploration of different directions for interpretation, without compromising the model fit. This property makes it possible to choose the solution that is easiest to interpret by optimising the simplicity of the component coefficients (*i.e.*, large absolute values for the entities that contribute more to a given component, and values close to zero for those entities with negligible contribution). This way, the interpretation can be performed more easily

by first identifying the entities with largest absolute coefficients in a given component, and then assigning a meaning that reflects what these entities have in common. For the present data, we rotate the initial solution in such a way that both the core and the component matrix **B** jointly tend to a simpler structure, using the joint orthomax rotation approach proposed by Kiers (1998). Then, we label each component based on the analysis of the largest absolute coefficients.

Table 4.4: Matricised (rotated) core array resulting from the application of a Tucker3 model of order $(4 \times 3 \times 4)$ to the undirected binary version of Van de Bunt's friendship networks. A1, A2, A3, and A4 refer to the first, second, third, and fourth components of matrix **A**. The notation used for B and C follows the same reasoning. The contribution of each combination of components to the model fit (82.65%) is given inside brackets. The largest entries are emphasised in bold.

	B1xC1	**B2xC1**	**B3xC1**	**B1xC2**	**B2xC2**	**B3xC2**	**B1xC3**	**B2xC3**	**B3xC3**	**B1xC4**	**B2xC4**	**B3xC4**
A1	0.24	2.29	0.34	−1.95	−0.02	−1.78	7.82	−2.47	0.18	−0.12	−0.01	−0.13
	(0%)	(0.6%)	(0%)	(0.4%)	(0%)	(0.3%)	(6.6%)	(0.7%)	(0%)	(0%)	(0%)	(0%)
A2	2.39	−1.64	−0.06	−1.86	−3.73	−0.91	1.05	−2.27	0.1	−1.19	−3.16	**−17.04**
	(0.6%)	(0.3%)	(0%)	(0.4%)	(1.5%)	(0.1%)	(0.1%)	(0.6%)	(0%)	(0.2%)	(1.1%)	*(31.4%)*
A3	−3.76	2.74	−0.48	6.22	−0.01	−0.68	−0.91	0.41	−0.93	0.04	−0.16	0.15
	(1.5%)	(0.8%)	(0%)	(4.2%)	(0%)	(0%)	(0.1%)	(0%)	(0.1%)	(0%)	(0%)	(0%)
A4	1.31	−0.02	2.04	0.6	**−19.68**	−1.88	0.28	0.1	−3.44	−0.51	−0.66	−6.24
	(0.2%)	(0%)	(0.4%)	(0%)	*(41.8%)*	(0.4%)	(0%)	(0%)	(1.3%)	(0%)	(0%)	(4.2%)

Before delving into the details of the interpretation process, it is important to understand the meaning of the model outputs. The obtained Tucker3 model summarises the set of entities pertaining to each mode through a few components (as given in **A**, **B**, and **C**) and concisely describes the relations between these components (as given in $\mathscr{G}$). The entries of the component matrices **A**, **B**, and **C** give the weights (also referred to as scores or coefficients) of the corresponding entities (nodes, SNA measures, and time steps, respectively) in a given level of a given mode. These component matrices have as many columns, or levels, as the number of components defined in the order of the estimated Tucker3 model, which is $P = 4$ for matrix **A**, $Q = 3$ for matrix **B**, and $R = 4$ for matrix **C**. The rotated component matrices obtained in our experiments are provided in Appendix A. In turn, the core tensor $\mathscr{G}$ contains the weights of all possible triads (combination of components, for the three modes) and can be understood as a strongly reduced version of the full data array in terms of the summarised entities of each mode (Kiers and Van Mechelen, 2001). According to Tucker (1966), each core element g_{pqr} is the score of a *subject type p* on a *latent variable q*, in a *prototype condition r*. These weights reflect the strength of the interaction between the corresponding components and provide information on their relative contribution to the total fit (or variation explained by the selected model). Thus, the largest weights signal the most salient patterns, *i.e*, those that better capture the internal structure of the original three-order tensor. For orthogonal models, like the one we use, the squared core entries divided by the total sum of the squared core entries are proportional to the variation explained by the corresponding combination

of components (Kiers and Van Mechelen, 2001). Therefore, in order to find the most representative combination of components, one must find the largest entry in the core. The rotated core array obtained for our case study, along with the contribution of each core entry to the model fit, are given in Table 4.4. The results show that the interaction of components with the highest contribution to the model fit, *i.e.*, the one that is more important for understanding the internal structure of the original data, is the interaction $(4,2,2)$, followed by the interaction $(2,3,4)$. The interaction associated with the core entry g_{422} explains 41.8% of the core variation and relates the fourth component of mode A ($\mathbf{A}_{:,4}$) with the second components of modes B and C ($\mathbf{B}_{:,2}$ and $\mathbf{C}_{:,2}$, respectively). Analogously, the core entry with the second highest magnitude (g_{234}) has a contribution of 31.4% to the model fit and associates the second component of mode A ($\mathbf{A}_{:,2}$) with the third component of mode B ($\mathbf{B}_{:,3}$) and the fourth component of mode C ($\mathbf{C}_{:,4}$). These two core entries alone explain 73.2% of the model fit and, thus, encode a substantial part of the information contained in the original three-order tensor. Note that the largest core tensor values are associated with both the second and the third components of matrix **B**. This suggests that a coordinate system incorporating these components of matrix **B** will be able to recreate a faithful structural image of the original data. Hence, the maximum-variance 2D space where the trajectories will be projected will be defined by the second (x-axis) and the third (y-axis) components of matrix **B**, so as to capture the maximum possible information contained in the raw data in a 2D space.

For the purpose of our analysis, matrix **B** is the most relevant component matrix since the common space where the spatio-temporal trajectories will be projected is defined by its components. Since our goal is to capture the maximum amount of information in a bidimensional space, we select the most representative components of matrix **B** as the axes of this space. Hence, the second component $\mathbf{B}_{:,2}$ defines the x-axis and the third component $\mathbf{B}_{:,3}$ defines the y-axis. Figure 4.10 depicts the coefficients of mode B's entities (*i.e.*, the node-level measures) in the space spanned by the selected components ($\mathbf{B}_{:,2}$ and $\mathbf{B}_{:,3}$). The labels of these axes are assigned indirectly by first identifying the entities that contributed most to their formation and then conceptualising what they jointly represent. The entities' contribution to each component is reflected in the magnitude of their coefficients, with larger absolute coefficients meaning larger contribution. Thus, we focus on the entities with largest absolute coefficients to label the components that define the x-axis and the y-axis. Following this procedure we deduce that the second component of matrix **B** (Figure 4.10) opposes the clustering coefficient ($b_{5,2} = -0.59$) with the betweenness centrality ($b_{4,2} = 0.58$) and degree centrality ($b_{1,2} = 0.53$) measures. The clustering coefficient measures the cohesiveness of the local group in which a student is embedded, being large values often associated with strong relationships within a small and very close group. In turn, the degree and betweenness measures are related to the number of relationships an individual maintains and the extent to which those relationships are established with people belonging to different groups. Large values of degree and betweenness centralities are indicative of high social activity and interaction with diverse social

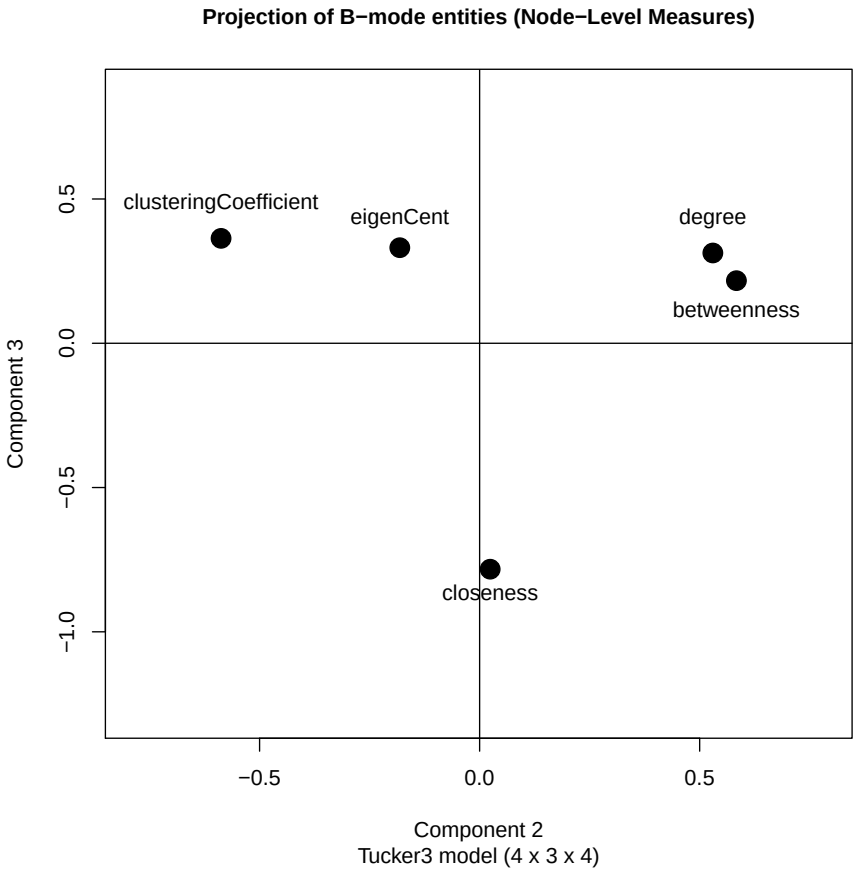


Figure 4.10: Projection of the coefficients of matrix **B** in the bidimensional space defined by the two most representative components of mode B, namely, the second and the third components. Mode B is associated to the column-entities (*i.e.*, the node-level measures) of the original three-way array χ.

groups. Hence, based on our interpretation, component $\mathbf{B}_{:,2}$ captures the dichotomy between social depth and social breadth and we label the associated x-axis as *social depth VS social breadth*. Since the clustering coefficient has a negative score, and the degree and betweenness centralities have both a positive score in this component, the left-hand side of the x-axis is associated with social depth, whereas the right-hand side is associated with social breadth. Thus, movements toward one of these directions signal a social behaviour that is more in line with the corresponding social dimension (social depth or social breadth). Regarding the third component of the same matrix, denoted by $\mathbf{B}_{:,3}$ and associated with the y-axis, we observe that this component is strongly related with a single measure: the closeness centrality (Figure 4.10). This is revealed by a large negative coefficient of $b_{3,3} = -0.78$. The remaining measures have scores lower than 0.4 and, therefore, their contribution to this component can be considered negligible. Since this component is strongly associated with the closeness centrality, we label the y-axis likewise. The negative sign of the closeness centrality coefficient indicates that large negative values in the associated y-axis correspond to high closeness centrality, whereas large positive values indicate low closeness centrality. Thus, students whose trajectories are moving toward the negative side of the y-axis are improving their ability to efficiently get in touch with others in the network.

Based on the above interpretation, we deduce that the most desirable social position in the 2D space, in terms of network centrality, is both the leftmost part of the third quadrant and the rightmost part of the fourth quadrant. Both are associated with high proximity to the rest of the network (high closeness centrality) but differ in terms of social behaviour: while in the former the presence of tight social grouping is more prevalent (large clustering coefficient), in the latter the social relationships are more expansive, in the sense they are established with a larger number of people (high degree centrality) belonging to different social groups (high betweenness centrality). Therefore, students whose trajectories take the direction of either the third or the fourth quadrants are improving their network centrality. Regarding the amplitude, if the coordinates of a given student's trajectory lie roughly in the same region, then it is assumed that the student's social position is stable. An analogous reasoning holds for the opposite scenario.

After decomposing the three-order tensor and assigning a label to the axes of the bidimensional space, we define the trajectories of each student following the procedure described in Section 4.4.2.

4.5.4 Analysis of Node-Level Spatio-Temporal Trajectories

As already stated by Van De Bunt et al. (1999), in this friendship network there is a clear transition between the moment where almost no student knows each other to the moment where most students are connected by friendship ties. This can be ascertained through visual inspection of the social networks depicted in Figure 4.8, where we observe a sharp increase in the network density from t_0 to t_1. The rise of network density reveals an intense period of socialisation and tie formation during the first weeks of university. In the following time steps, the density of the network stabilises, since

the friendship groups start to settle down. This phenomenon is also corroborated by the position of the students' trajectories in the bidimensional space, which are illustrated in Figure 4.11. By projecting the trajectories of all students in a common 2D space, we observe that, despite the clutter created by their spatial proximity and overlay, the origin of most of these trajectories lies in the origin of the space. Such coordinates are indicative of the initial lack of connections among students. Nevertheless, there are a few exceptions, namely, students 2 and 16, and students 17 and 31, whose friendship emerged in a former social context. The origin of these students' trajectories lies in the lower part of the third quadrant, thus emphasising their early advantage in terms of network centrality. Another conclusion we can draw from the visual analysis is that all trajectories lie either in the second or in the third quadrant of the 2D space. This indicates that the socialisation among freshmen evolves in the direction of social depth instead of social breadth. In fact, in the analysed data, there are no examples of students whose trajectories take the path of social breadth. This result makes sense, as we are analysing the formation and growth of a friendship network among students who did not know each other when the data were first collected, but who shared the same physical space and performed similar activities in the following time steps. The fact these students became classmates and spent most of their days in the same social environment increased the opportunities for repeated interaction and the establishment of interpersonal bonds, thus creating a fertile ground for close friendships to bloom and develop over time.

Since one of the goals of our case study is to illustrate the proposed methodology, here we focus only on the analysis of two spatio-temporal trajectories, namely, the trajectories of students 8 and 16. The reasons for selecting these two specific students are the following: (i) availability of complete data for both students since they answered the seven questionnaires (Table 4.1); (ii) they are well represented in the model since their corresponding fit is large (89.46% for student 8 and 94.73% for student 16), as can be ascertained from Table A.3 in Appendix A; (iii) they stand at two distinct social positions at the first time step t_0, since one of them has a connection to another student while the other does not have any initial connection; (iv) they belong to the same dynamic community but take different paths when the community splits in the last time interval $[t_5, t_6]$. Overall, these two students display a distinctive behaviour over time that represents well the social patterns found in the whole set of students.

Spatio-Temporal Trajectory of Student 16

The first spatio-temporal trajectory to be analysed is the one pertaining to student 16, who is a non-smoker female attending the 4-year program. Her trajectory is depicted in Figure 4.12. The trajectory is somewhat irregular, showing abrupt changes in direction during the first time steps (*e.g.*, in $[t_0, t_1]$, $[t_1, t_2]$, and $[t_2, t_3]$). However, after t_3 the trajectory stabilises and starts exhibiting a leftwards trend. This trend reveals both the emergence of a close group of friends, whose bonds strengthen over time, and a social behaviour denoting a preference for social depth. From the

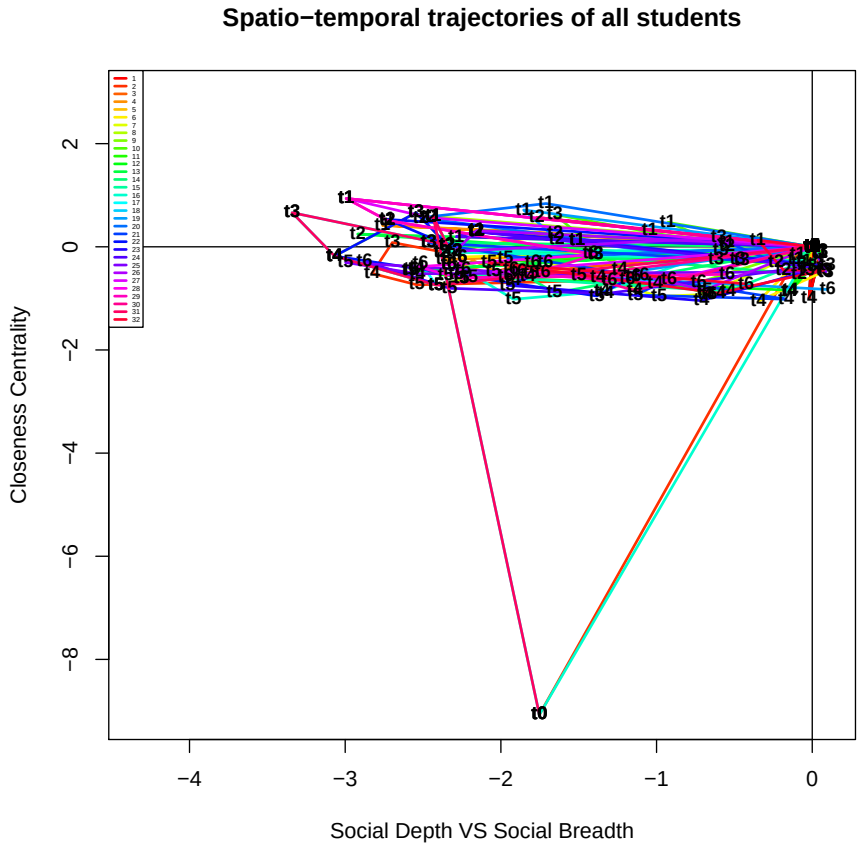


Figure 4.11: Trajectories of all students in the space spanned by the two most representative components of matrix **B**.

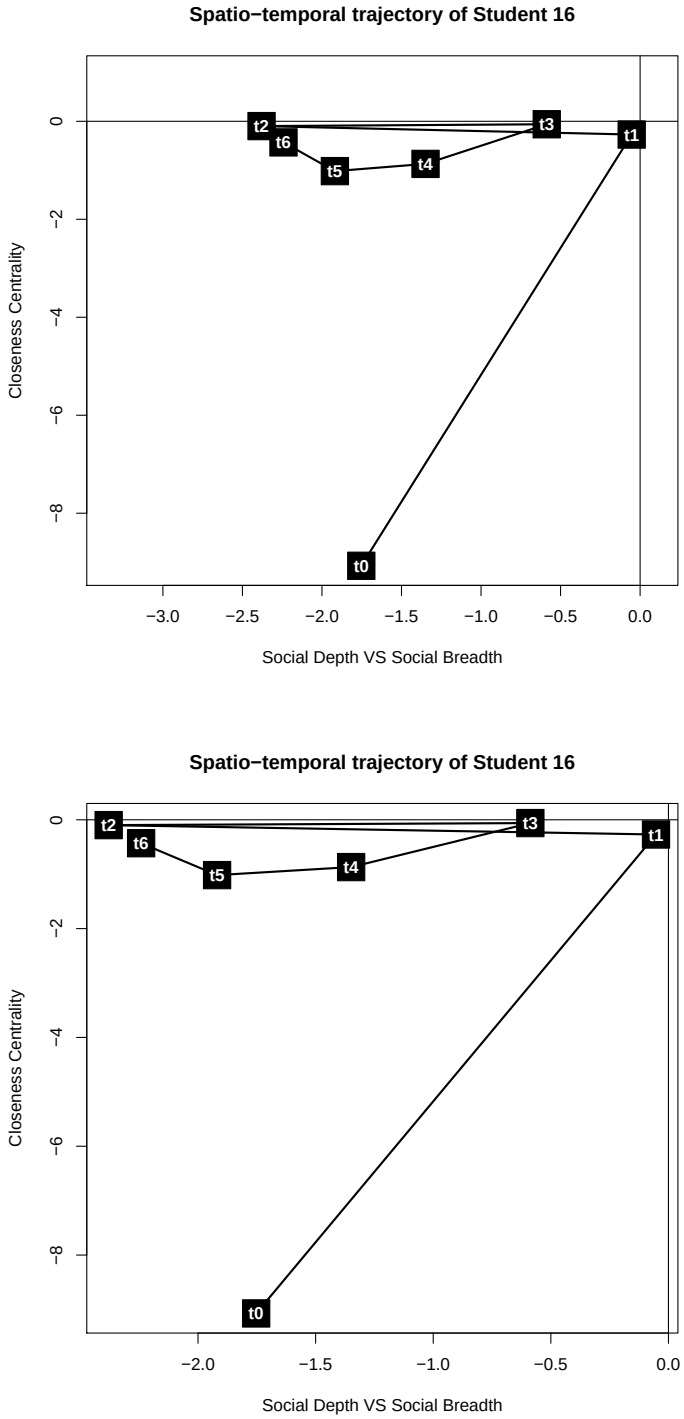


Figure 4.12: Spatio-temporal trajectory of student 16 in the common 2D space defined by the two components of matrix **B**. The bottom figure provides a closer look at the trajectory.

analysis of Figure 4.12, we also conclude that the position of the complete trajectory lies mostly in the third quadrant of the 2D space, and this position is consistent throughout the whole period of observation. Concerning amplitude, we observe that the relative distance between two of the farthest points in the trajectory, namely, t_0 and t_1, is large compared to, for instance, the one observed for student 8's trajectory (Figure 4.13)[1]. Another fact we can draw from the analysis is that student 16 is one of the few freshmen who had friendly ties with another student from the set in the first time step, as can be deduced from the position of t_0 in the bottom side of the third quadrant. In terms of coordinates, this represents the best social position that student 16 achieved during the time span under analysis, as indicated by the large absolute scores in both axes. While most students start their trajectories at the origin of the plane, student 16 begins with some social advantage, which is reflected in a high closeness centrality (negative y-value) and a high social depth (negative x-value) at t_0. However, this advantage is lost in the next time step as the trajectory takes the direction of the origin of the space. Since closeness centrality is a global measure, in the sense it is based on the computation of shortest paths among nodes in a network, the decrease in the closeness centrality from t_0 to t_1, denoted by the upward movement of student 16's trajectory, signals a global change in the friendship network. This complies with the background knowledge, as this time interval corresponds to an intense socialisation period among freshmen students during the first weeks of university. With the formation of friendships during these weeks, students who were once isolated in the network at t_0 become connected at t_1. As a consequence, the high network centrality conferred by the connections of student 16 in the first time step decreases as a result of this socialisation phenomenon. In the remaining time steps, the closeness centrality remains relatively constant and close to zero. The displacement of the trajectory in the x-axis from t_1 to t_2 suggests that the social interactions of student 16 were limited to a close group of friends. However, in the following time step we observe a similar horizontal shift of the trajectory but in the opposite direction. This movement can be interpreted as a broadening of the circle of friendships of student 16 during $[t_2, t_3]$, revealed by a social behaviour that is more in line with social breadth than with social depth. Yet, this behaviour is not constant over time since the trajectory retakes its path toward social depth in the subsequent time steps, suggesting an extension of the initial friendship group into a larger but still cohesive group of students.

The previous analysis suggests that the social behaviour of student 16 over time denotes a preference for lasting and intimate relationships with a close group of friends. During the first weeks, this student interacted mostly with former friends and only after this initial period she intensified her social activity with other students. This socialisation process led to the enlargement of her previous group of friends.

[1] The visual comparison of the amplitudes should be based on the top subfigures, since they map the trajectories in a scaled 2D space

Spatio-Temporal Trajectory of Student 8

The second spatio-temporal trajectory to be analysed is associated with a smoker female also attending the 4-year program, designated as student 8. As can be ascertained from Figure 4.13, this student's trajectory is compact when compared to the trajectory of student 16, and is stable regarding its overall position in the plane since all time steps are located roughly in the intersection between the second and third quadrants, with the exception of the first time step t_0. However, when taking a closer look at the dynamics of this student, it is apparent some inconstancy and lack of trend in the trajectory's direction, namely in terms of closeness centrality. The majority of time steps are concentrated in the left side of the x-axis, which reveals a steady social behaviour of this student toward social depth. Although at the beginning (t_0) student 8 did not have any connections with the freshmen belonging to the analysed group, as indicated by the coordinates of the first time step, during the first weeks she was able to develop friendly ties with a small group of students. Her membership to this group remained stable over time and the large negative x-values suggest that this is a tightly-knit and cohesive social circle, where everyone is friends with each other. The observed oscillations in the trajectory's closeness centrality are related to global changes in the network structure. These changes are triggered by events such as the formation of new ties between previously unconnected students, or the dissolution of ties. The densification of the network resulting from the socialisation among students shortens the network paths between them, thereby positively influencing the average closeness centrality of the network. However, some students have higher closeness centrality values than others as a consequence of being positioned in a more central region of the network. Given this, we can deduce that during the time interval $[t_0, t_3]$ student 8 was located in a peripheral region of the network (time steps t_0, t_1, t_2, and t_3 have positive y-values), and only after t_4 this position becomes more central and, therefore, closer to the remaining students in the network (time steps t_4, t_5, and t_6 have negative y-values).

From the previous analysis, we deduce that, despite the lack of initial connections, student 8 was able to quickly integrate a social group and develop close friendships with its members. These friendships persisted over time and remained quite stable, thus indicating a social behaviour inclined to social depth.

4.5.5 Analysis of Community-Level Spatio-Temporal Trajectories

The proposed methodology allows the exploration of an evolving network at two levels of analysis, the node-level and the community-level, using spatio-temporal trajectories. In the previous subsection, we gave examples on how to analyse and interpret spatio-temporal trajectories of single nodes, in order to grasp insight into their temporal evolution and detect local and global changes. However, the analysis of the evolution of single nodes may not suit the purpose of several applications (*e.g.*, temporal analysis of customer segments or research fields), where the goal is to discover and

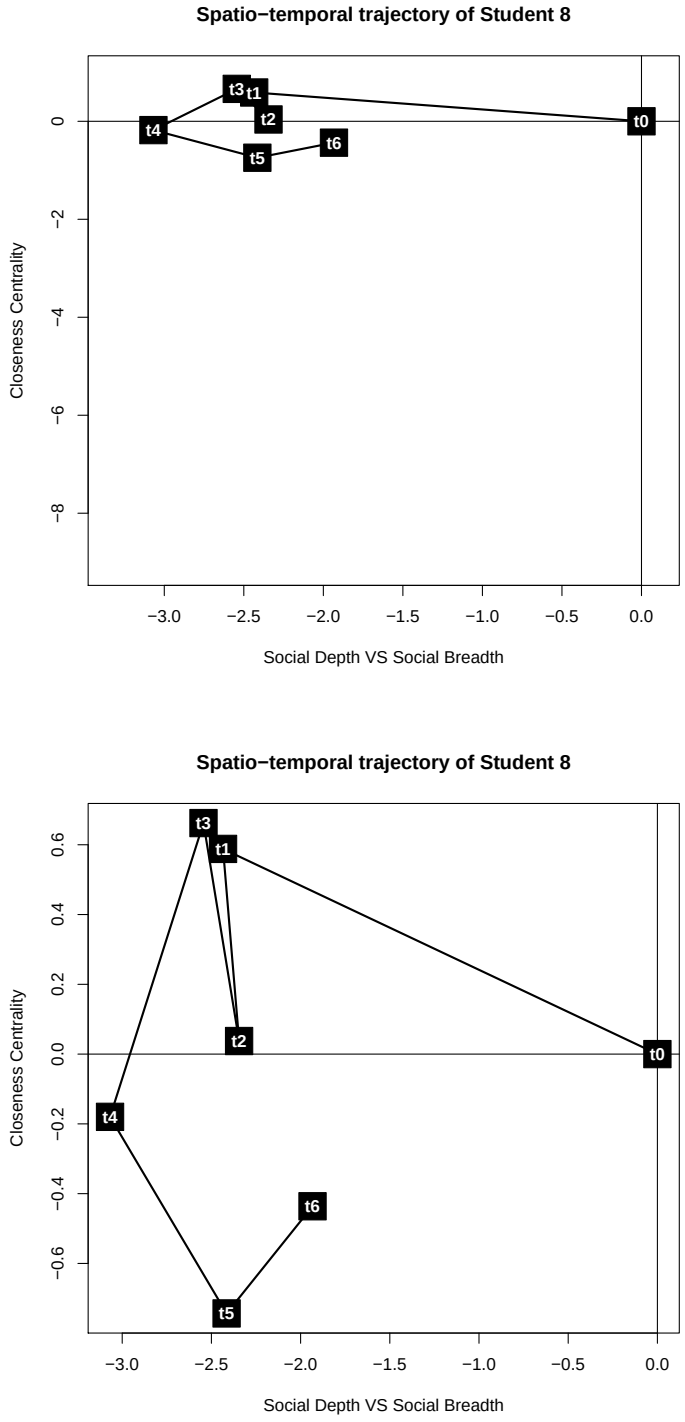


Figure 4.13: Spatio-temporal trajectory of student 8 in the common 2D space defined by the two components of matrix **B**. The bottom figure provides a closer look at the trajectory.

understand temporal patterns in higher-level structures of the network. Besides, the representation of all node-level trajectories in a common 2D space can become quite cluttered, as apparent in Figure 4.11, especially when dealing with large networks. An alternative (or complementary) approach is to abstract the network into communities, *i.e.*, densely connected groups of nodes, and examine their evolution. In this section we introduce this approach, by changing the level of analysis from the node to the community, and focus on the analysis of a trajectory that depicts the dynamics of such cohesive sets of nodes.

In the context of this case study, a community is defined as a cohesive set of students who share friendship ties and, thus, interact closely and frequently with each other (*e.g.*, group of friends). Such communities tend to evolve gradually as new students join, old members leave, and the relationship among members changes over time. In order to generate trajectories of dynamic communities we carried out the steps described in Section 4.4.2.

Table 4.5: Number of communities (with and without singletons) and modularity values associated with the network partitions discovered by the Louvain method. The Louvain method was independently applied to each network snapshot of Van de Bunt's temporal friendship network.

	t_0	t_1	t_2	t_3	t_4	t_5	t_6
Number of Communities	30	12	13	11	11	10	14
Number of Communities (without singletons)	2	5	4	5	4	4	6
Average Size of Communities (without singletons)	2	5	5.75	5.2	6.25	6.5	4
Modularity	0.5	0.67	0.58	0.57	0.49	0.37	0.48

Communities were extracted independently at each time step using the Louvain method. The community detection results are summarised in Table 4.5. For all time steps, the number of singletons, *i.e.*, isolated students in the network, is large and higher than the number of communities with two or more members. These singletons correspond either to students who developed friendships outside this specific group of freshmen, or to students who did not answered the questionnaire at a given time step. The average size of communities tends to increase over time, as would be expected from the intensification of the socialisation among students. Still, we observe a drop in the communities' size at time step t_6, which is partly explained by critical evolutionary events in the communities' life-cycles, namely, community splits. Concerning the quality of the network partitions, based on the analysis of the modularity values, we conclude that the communities discovered by the Louvain method are meaningful for all time steps and, thus, we can proceed with the analysis. The next step involved the application of the MECnet event-based framework to the set of previously discovered communities with the aim of identifying dynamic communities and obtain their temporal trajectories. We set the user-defined thresholds of MECnet to $\tau = 0.5$ (matching threshold) and $\lambda = 0.2$ (split threshold). These values are less conservative in order

to enable the detection of persistent communities. Since the first time step t_0 is atypical, as the corresponding network snapshot is very sparse and splintered into many singletons, we assume that dynamic communities are a result of merges of nodes in $[t_0, t_1]$ and only take form in t_1. Having information about the dynamic communities, we proceeded with the aggregation of the SNA measures as explained in Section 4.4.2, and further computed the coordinates that define the spatio-temporal trajectories of the dynamic communities. For illustrative purposes, we select the most persistent community over time, which is also one of the largest communities from the set of dynamic communities of students. Henceforth, we designate this community as *dynamic community A*.

Spatio-Temporal Trajectory of Dynamic Community A

Table 4.6: Membership of dynamic community A, given in terms of students' IDs, for each time interval. Note that in the last time interval this dynamic community splits into two communities: $Com^1_{t_6}$ and $Com^5_{t_6}$.

Time Intervals	Community Instance	Community Membership
$[t_0, t_1]$	$Com^1_{t_1}$	$\{2, 8, 10, 16, 17, 22, 31\}$
$[t_1, t_2]$	$Com^2_{t_2}$	$\{2, 8, 10, 16, 17, 22, 31\}$
$[t_2, t_3]$	$Com^4_{t_3}$	$\{8, 10, 17, 22, 31\}$
$[t_3, t_4]$	$Com^2_{t_4}$	$\{2, 8, 10, 16, 17, 22, 31\}$
$[t_4, t_5]$	$Com^2_{t_5}$	$\{2, 8, 10, 16, 17, 22, 31\}$
$[t_5, t_6]$	$Com^1_{t_6}$	$\{2, 3, 7, 10, 16, 23, 24\}$
$[t_5, t_6]$	$Com^5_{t_6}$	$\{8, 17, 22\}$

Dynamic community A, denoted as Com^A, comprises an average of six students and includes students 8 and 16. The detailed membership is given in Table 4.6. It is one of the largest dynamic communities in the temporal friendship network and the most persistent one. The stability of Com^A manifests itself in the high number of consecutive survivals observed in its life-cycle: $LC_{Com^A} = \{merge_{[t_0,t_1]}, survival_{[t_1,t_2]}, survival_{[t_2,t_3]}, survival_{[t_3,t_4]}, survival_{[t_4,t_5]}, split_{[t_5,t_6]}\}$. This life-cycle tell us that dynamic community A was created in t_1 as the outcome of a merge of nodes during $[t_0, t_1]$, survived until t_5 by keeping most of its members, and split into two communities in the last time interval $[t_5, t_6]$. Its temporal trajectory is defined as $TT_{Com^A} = \{Com^1_{t_0}, Com^2_{t_0}, Com^9_{t_0}, Com^{11}_{t_0}, Com^{21}_{t_0}, Com^1_{t_1}, Com^2_{t_2}, Com^4_{t_3}, Com^2_{t_4}, Com^2_{t_5}, Com^1_{t_6}, Com^5_{t_6}$. This temporal trajectory is merely descriptive and fails in providing a meaning to the critical evolutionary events highlighted in the community's life-cycle. We attempt to overcome this problem by assigning a spatial context to the temporal trajectory, using as a basis the structural patterns extracted by the Tucker3 model. To achieve this goal, we project the temporal trajectory of dynamic community Com^A in the space spanned by the two most representative components of matrix **B**, which is exactly the same 2D space used for projecting the node-level trajectories. The coordinates of the community-level

trajectory are obtained by averaging the coordinates of the nodes pertaining to each community instance of the dynamic community. This way, we are able to generate the spatio-temporal trajectory of Com^A, which is depicted in Figure 4.14.

The analysis of Figure 4.14 conveys information on several aspects of the evolution of community Com^A. The life-cycle of this community is represented in the spatio-temporal trajectory and we can graphically see the merge of nodes that enabled the emergence of Com^A, the subsequent survivals, as well as the split that occurred in the last time interval. This dynamic community takes form in t_1 as a result of the development of friendly ties between seven students (see Table 4.6). Some of these students were already connected to someone in the group, while others were isolated in the network at t_0. Each one of these situations is reflected in the community's trajectory, as there are two different paths converging into point t_1: one beginning at the origin of the space and another at the bottom of the third quadrant. The trajectory also indicates that dynamic community A occupies a peripheral position in the network at time steps t_1, t_2, and t_3, as suggested by the low closeness centrality (higher y-values). The leftward movement observed during these three time steps also suggests a strengthening of the friendship bonds among the members of the community (lower x-values). However, there is a turning point in the evolution of community A at t_4, which is revealed by a shift of the trajectory toward the right side of the third quadrant. From t_4 onwards, the rightward movement of the trajectory indicates an improvement of the community's centrality both in terms of proximity to other communities (lower y-values) and connectivity in the network (higher x-values). This change in centrality is related to an increase in the breadth of connections of a few community members, achieved through the formation of new ties with students from other communities. Note that the dynamic behaviour of community A until t_5 reflects a general pattern in the social evolution of its constituent members. Since the temporal pattern described thus far is shared by all members pertaining to Com^A, the dynamic community survives. However, this is not always the case, as revealed by the split of Com^A into two communities in the last time interval $[t_5, t_6]$. This community split is explained by slight changes in the closeness centrality of two subgroups of members, as can be ascertained from Figure 4.14, and might be a consequence of the formation of new ties during the previous time interval. The split event experienced by Com^A at t_6 gave rise to two distinct communities: one comprised of three students and another comprised of seven students (see Table 4.6). The smaller community consists of persistent members of dynamic community A, namely, students 8, 17 and 22, whereas the larger community comprises old members (students 2, 10 and 16) as well as new members (students 3, 7, 23 and 24). This suggests that students 2, 10, and 16 were responsible for the improvement of the community centrality during $[t_4, t_5]$, and this improvement was triggered by their social interaction with students outside the community (*e.g.*, students 3, 7, 23 and 24). This interaction enabled the development of new friendships with these new students and contributed to the emergence of a new community. This new community has a network position that is slightly more central than the one occupied by

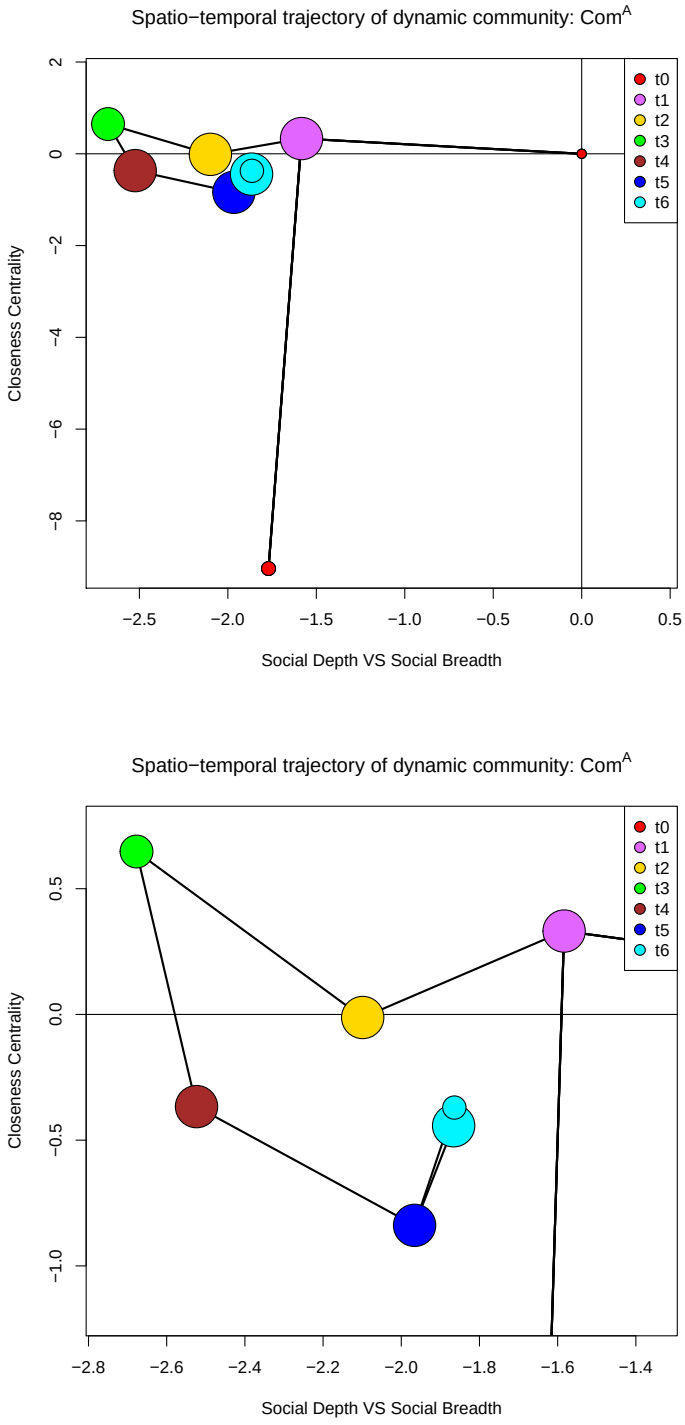


Figure 4.14: Spatio-temporal trajectory of dynamic community A in the common 2D space defined by the second and the third components of matrix **B**. The bottom figure provides a closer look at the trajectory. In both figures, the size of the points is proportional to the size of the associated community instances (larger points indicate a larger number of members).

the smaller community, as indicated by their coordinates in the y-axis of the 2D space.

The previous analysis of community A's trajectory allowed us to identify two distinct trends in its evolution: a leftward trend starting at t_1 and ending at t_3, and a rightward trend starting at t_4 and ending at t_6. In the former (leftward trend), the trajectory takes the direction of social depth and lies in the second quadrant, which is associated with low closeness centrality. This trend signals a period characterised by little interaction with students from other communities and by an increase of the community's internal cohesiveness. In the latter (rightward trend) the trajectory reverses its direction toward social breadth and higher closeness centrality, thus marking a period of outward expansion of the community to the remaining network and consequent improvement of its centrality. This shift ultimately culminated in the split of the community into two distinct communities.

4.6 Conclusions and Future Work

Visualisation has been a central topic in SNA since its inception. Node-link diagrams, or graphs, are the standard and most popular model for representing and visualising static networks. However, they have inherent limitations when applied to dynamic networks. This calls for the development of novel visualisation methods tailored to the specificities of evolving networks. In this chapter, we tackle this problem by introducing a new approach to model and visualise dynamic social networks. The approach is based on the extraction of spatio-temporal trajectories from tensorial representations of time-evolving social networks. The projection of trajectories of single network entities in a concise 2D space helps understand the evolution of these entities with respect to latent properties of the underlying dynamic social network. The use of the Tucker3 model to generate a common 2D space guarantees that this space is a faithful representation of the main structure of the three-way raw data and ensures that the displayed patterns are free of noise. A case study using a set of self-reported friendship networks among university freshmen revealed that the proposed approach has several desirable features, being concomitantly simple, informative, and compact, thus allowing the visual inspection and understanding of structural changes and patterns in the evolution of entities in dynamic social networks. We also showed that tensor modelling and analysis can be successfully used to provide high-order summarisations of both the structure and behaviour of evolving social networks, at different levels of analysis.

The proposed methodology has broader uses than the one presented here and potential application in several real-world business tasks, such as churn prediction and viral marketing. In the former application the goal is to predict the probability of customer churn[2] in telecommunications companies through the analysis of networks of phone calls between subscribers (*i.e*, customers) over time. In this context, our methodology could be used to first map the trajectory of each

[2] *Customer churn* refers to the propensity of customers to left a company, usually in favour of a competitor, in a given time period.

customer appearing in the "who-calls-whom" network, and then comparing the distance between the trajectories of current customers with the *churning trajectories* (*i.e.*, trajectories of customers who have left the company). Churn prediction is a topic with inherent commercial interest and practical value for companies. The timely detection of a prospective churner can help companies improve the efficiency of their marketing plans and develop targeted actions to avoid customer loss. Besides, it can contribute to increase the profits, since it is generally believed that it is more expensive to attract new customers than to retain existing ones. In the later application, the main goal is to analyse the existing social network of customers, find those with strongest influence in the market, and then use them as "seeds" for promoting a product or service, or for increasing brand awareness and recognition. This idea is based on the assumption that product adoption spreads from customer to customer through social networks. By primarily marketing customers with high network value, companies can take advantage of word-of-mouth advertising, by leveraging customers themselves to carry out the most promotional effort and act as product advocates. Marketing a small set of customers, who are expected to positively influence the probability of their immediate connections in buying the product, can prove to be more cost effective than traditional marketing methods. In fact, according to the findings of Hill et al. (2006), the probability of product adoption is higher if the product recommendation is made by network neighbours, *i.e.*, by people who are directly connected to the prospective customer in a social network. However, the customer selection process in viral marketing is usually based on a single snapshot of their social networks. Using our methodology, we could add temporal considerations to this process, by deriving the network value of a customer based on his/her spatio-temporal trajectory. This way we would expect to improve the success rate of word-of-mouth.

Although our methodology has several business applications and advantages over the state-of-the-art visualisation approaches, it is not free from limitations. The first identified limitations are related to the scope of application. One of the requirements for the application of a Tucker3 model is the presence of three-way interactions, or at least two non-negligible two-way interactions, in the input three-order tensor. Since our approach strongly relies on the Tucker3 model, if the data at hand do not meet this requirement the methodology, as we devised it, cannot be applied. Another drawback is the non-incremental nature of the methodology and the consequent need to traverse most of its steps as new network data are available. This shortcoming is explained by the nature and purpose of the methodology, which is tailored for performing retrospective descriptive analysis of an evolving network based on historical data. As a consequence, we also assume that the number of nodes is known *a priori*, in order to ensure that the dimensionality of the three-order tensor is fixed. However, in dynamic settings, new nodes can be added to, or removed from, the network and thus it is important to allow the dimensionality of the tensor to change over time. A possible solution to address this challenge is to use PARAFAC2 (Harshman, 1972; Chew et al., 2007) for irregular tensors instead of a Tucker3 model. Concerning communities, our visualisation approach

at the community-level inherits some of the limitations of the MECnet event-based framework thus being circumscribed to the analysis of crisp communities. The second limitation is related to the visual and computational scalability. In the case study, we analysed a very small network, which does not reflect the size of the networks available today. Processing large dynamic networks can be demanding when using tensor decomposition models, due to their high computational complexity. On the other hand, performing network evolution analysis over long time spans can pose problems to the readability of spatio-temporal trajectories in the low-dimensional Tucker space, due to the risk of visual cluttering. The third limitation of this research work is the absence of a user study to evaluate the proposed visualisation approach and decide whether it is helpful and useful. The identified limitations and problems require further investigation and offer interesting directions for future research.

Part III

Multicriteria Decision Analysis for Prescriptive Analytics

Chapter 5

Multicriteria Decision Analysis: A Case Study in the Paint Shop of an Automobile Assembly Plant

The painting activity is one of the most complex and important activities in the automobile manufacturing. The inherent complexity of the painting activity and the frequent need for repainting usually turn the painting process into a bottleneck of automobile assembly plants, which reflect in higher operating costs and longer overall cycle times. One possible approach for optimising the performance of the paint shop is to improve the efficiency of the colour planning. This can be accomplished by evaluating the relative merits of a set of vehicle painting plans. Since this problem has a multicriteria nature, we resort to the MCDA methodology to tackle it. A recent trend in the MCDA field is the development of hybrid approaches, which are used to achieve operational synergies between different methods. Here we apply, for the first time, an integrated approach that combines the strengths of the AHP and the PROMETHEE, aided by Geometrical Analysis for Interactive Aid (GAIA), to the problem of assessing vehicle painting plans. The management of the assembly plant found the results of value and is currently using them in order to schedule the painting activities such that an enhancement of the operational efficiency of the paint shop is obtained. This efficiency gain has allowed to bid for a new automobile model to be assembled at this specific plant.

5.1 Introduction

In recent years, the increasing competitiveness of the global market, as well as the burst of the so-called Global Financial Crisis (also known as the 2008 financial crisis), forced companies to rethink their processes in order to raise the levels of efficiency, responsiveness, and flexibility. In

such contexts, resorting to MCDA to assist in both strategic and operational decision problems can be a decisive step toward achieving these goals. MCDA is a structured approach to determine preferences among possible options that achieve more than one objective. Both the academic attention devoted to the field of MCDA and the widespread application of its methods in real-world decision problems are a reflection of the prominence and advantages of MCDA approaches in aiding and supporting decision-making.

Over the years several MCDA methods have been proposed (Goodwin and Wright, 2004; Turskis and Zavadskas, 2011) and this number is ever-increasing. Two of the most popular are the AHP (Saaty, 1977, 1986, 1990) and the PROMETHEE (Brans, 1982; Brans and Vincke, 1985; Brans and Mareschal, 1994). The former pertains to the normative (or American) school of thought, which is represented by methods that perform preference aggregation through value functions, such as Multiple Attribute Utility Theory (MAUT), Multiple Attribute Value Theory (MAVT), Analytic Network Process (ANP), and Simple Multi-Attribute Rating Technique (SMART). The latter belongs to the European (or French) school of thought, whose theoretical underpinnings rely on the concept of outranking. The families of methods influenced by this school, such as ELimination and (Et) Choice Translating REality (ELECTRE) and PROMETHEE, perform outranking so as to eliminate options that are dominated in a particular sense. Despite the popularity of AHP and PROMETHEE, none of these methods is better than the other, with both having strengths and weaknesses.

Realising that there is a lack of a universally recognised best method for MCDA that would suit every type of multicriteria decision-making problem and is impervious to weaknesses (Chang et al., 2013), researchers started to explore synergies between well established methods in an attempt to boost the strengths and mitigate the weaknesses associated with each individual method. As a result, hybrid approaches combining within a single framework two or more MCDA methods, or one MCDA method with other methodologies, have emerged as a trend in the field (Keune et al., 2013). Examples of hybrid approaches include, among others, integrated AHP approaches (with, for instance, Genetic Programming, PROMETHEE, and Data Envelopment Analysis) (Badrí, 2001; Macharis et al., 2004; Wang et al., 2008), integrated ANP approaches (with, for instance, Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), Decision-making Trial and Evaluation Laboratory, and Multi-Objective Programming) (Shyur and Shih, 2006; Yang et al., 2008; Demirtas and Üstün, 2008), and integrated fuzzy approaches (for example, fuzzy ANP with fuzzy TOPSIS, and fuzzy ELECTRE) (Kabak et al., 2012). The research presented in this chapter follows this trend and proposes the use of a hybrid approach that integrates the AHP and the PROMETHEE to address a multicriteria decision problem in the paint shop of an automobile assembly plant.

The automobile industry has been one of hardest hit by the 2008 financial crisis, which led to a sharp fall in industry sales. One of the most complex and important activities in automobile

manufacturing is the painting activity (Geffen and Rothenberg, 2000; Li et al., 2007). According to Leichtling (2002), colour is an important marketing tool due to the psychological effects it can trigger and its intrinsic power in attracting (or repulsing) prospective customers. Given the influence that colour has on customers' purchasing behaviour, automobile companies devote great efforts to ensure the quality and attractiveness of their vehicles' colour appearance. However, the inherent complexity of the painting activity and the frequent need for repainting usually turn the painting activity into a bottleneck of automobile assembly plants. Thus, there is a great need for improving the operational efficiency of the paint shop.

To the best of our knowledge, previous research on paint shops (Ulgen and Gunal, 1998; Chung et al., 2001; Meunier and Neveu, 2012) focus on a particular aspect of the problem (the number of colour changes in the incoming sequence of vehicles) and the solution involves optimising a single objective function (minimizing the length of the schedule for the given time horizon, which is equivalent to minimizing the number of colour changes), based on a mono-criterion approach. For instance, Meunier and Neveu (2012) tackled this problem from a theoretical point of view, and resorted to linear programming and local search algorithms to find the sequence of cars in a paint booth that minimizes colour changes. The problem of vehicle sequencing in automobile paint shops has also been tackled from the perspective of energy consumption. For instance, Wang et al. (2009) proposed an optimal scheduling procedure for vehicle sequencing, which reduces the energy consumption of the paint shop. Their approach was based on selecting appropriate batch and sequence policies in order to reduce the energy consumption by means of the improvement of the paint quality and the decrease of the number of repaints. This chapter takes a different view on the analysis of paint shop problems by focusing on the problem of evaluating the relative merits of different painting plans. A painting plan can be briefly defined as a combination of vehicle types (single or mixed) with the number of distinct colours used to paint those vehicles in a given day. This concept is detailed in Section 5.4. The problem of evaluating painting plans differs from the paint shop scheduling problem in a number of ways. First, instead of considering a single operational feature (*e.g.*, number of colour changes, energy consumption), attributes of the entire painting plan are considered based on what is regarded as relevant by the management of the company (for instance, the vehicle types and the number of distinct colours used in a given day). Secondly, this work focuses on providing guidance to the Decision Maker (DM) in discovering a set of compromise solutions, rather than a single optimal solution, by treating the problem as a multicriteria decision problem. This implies the existence of a predetermined and finite number of options (the painting plans defined by the decision maker), which are subjectively appraised by the DM in terms of the degree to which they meet a set of multiple, and often conflicting, criteria. In fact, several human decision problems have a multicriteria nature and should be treated as such. The evaluation of painting plans would be simpler if the assessment was solely based on one criterion (for instance, reduction of paint consumption levels). However,

the appropriate treatment of this problem should embrace the inherent complexity of the painting system by taking into account all the relevant criteria affecting the evaluation. These criteria can be of qualitative or quantitative nature, and they reflect different dimensions of the evaluation process (*e.g.*, technological, economical, and environmental dimensions), which influence the decision-making process in various degrees. Besides, they are often conflicting since there exists no option optimising all the criteria at the same time. The conflicting nature of criteria requires a trade-off between them. The problem of evaluating painting plans is characterised by all the aforementioned features thus having a multicriteria nature. Due to the specific and complex nature of this problem, traditional mono-criterion or optimisation approaches are not appropriate as they do not perform a multicriteria evaluation of the values, do not allow for compromise, and are unable to handle the subjectivity, uncertainty, and presence of imperfect information in decision contexts. For these reasons, we resort to the MCDA methodology to address this problem.

In this chapter, we tackle the paint shop problem of evaluating different painting plans in automobile assembly plants using a hybrid methodological approach: the AHP-PROMETHEE. The integration of two MCDA methods makes this approach more elaborated from the methodological point of view. However, the methodology is comprised of simpler components that are easier to understand and work with, when compared with the stand-alone versions of the AHP and the PROMETHEE. In addition, this approach helps to simplify and structure the DM's input, thus leading to a higher commitment of the DM with the decision-making process. It is of utter importance for the DM to obtain robust and reliable results since these will be used to schedule the painting plans and, thus, the production of the assembly plant, in order to improve its efficiency. To achieve this goal, we propose the combination of the AHP and PROMETHEE by using them at different stages of the MCDA methodology. Although these methods pertain to different school of thoughts and, thus, are different in nature, operational synergies can arise from the integration of both methods, as already demonstrated in the literature (Macharis et al., 2004; Wang and Yang, 2007; Dağdeviren, 2008; Turcksin et al., 2011; Bogdanovic et al., 2012; Venkatesan and Kumanan, 2012; Gervásio and Simões da Silva, 2012; Agha et al., 2012; Avikal et al., 2013; Nasiri et al., 2013; Bansal and Kumar, 2013; Herva and Roca, 2013).

The contributions of this research are threefold. We introduce a novel multicriteria paint shop problem, namely, the evaluation of vehicle painting plans in automobile paint shops. We are also the first to apply the AHP-PROMETHEE integrated approach to this multicriteria problem, with the purpose of sorting the painting plans according to their ability to achieve the DM's goals. Thirdly, on a different level, we encouraged the adoption of a more systematic and rigorous approach in problem solving by the DM of the assembly plant, since previously his decisions have been mostly made based on business experience, technical knowledge, and analysis of non-integrated data. It is also expected that the outcomes of this research contribute to the improvement of the operational efficiency of the painting activity, this way releasing production capacity that together with the

one already released by the decrease in production (due to the economic crisis) allows the plant to apply for the production of a new automobile model. This constitutes the major motivation for conducting this study.

The chapter proceeds as follows. We begin by providing an overview of the historical background of MCDA. In the same section, we also introduce basic concepts of MCDA and the adopted terminology. Section 5.3 outlines the typical steps involved in the deployment of a MCDA procedure and describes the AHP, PROMETHEE, and AHP-PROMETHEE approaches. Section 5.4 provides a detailed description of the application of AHP-PROMETHEE to a real-world decision problem in the paint shop of one of Toyota's assembly plants. Section 5.5 summarises the chapter and offers some concluding remarks.

5.2 Historical Background and Concepts of MCDA

Decision analysis, decision aid or decision support, is a broad research field devoted to the development of theories and methodologies to address complex decision problems and their application to real-world decision contexts. The general aim of decision analysis is to support decision makers in the process of making sound and reasonable decisions, by means of structured and systematic procedures. A decision problem can be defined "by the options among which one must choose, the possible outcomes or consequences of these options, and the contingencies or conditional probabilities that relate outcomes to acts" (Tversky and Kahneman, 1981, p. 453). Many of the decision problems faced by decision makers contain multiple objectives that conflict with one another. The MCDA approach was devised to handle this type of problems.

MCDA is a problem-solving methodology to formally analyse decision problems where more than one point of view must be considered. It organises and synthesizes the information regarding a given decision problem in a way that provides the DM with a coherent overall view of the problem, while helping him/her advance towards a solution(s). MCDA is intended to support the DM's learning in what regards the nature of the multicriteria problem, his/her priorities, preferences, system of values, and objectives, so as to help him/her improve the quality of his/her decisions by making more rational and explicit choices. As the name implies, the decision problems falling within the scope of MCDA have a multicriteria nature, *i.e.*, they require the consideration of multiple criteria when comparing a set of options. Therefore, the concept of "optimal solution", which lies at the heart of mono-criterion or optimisation approaches, typically does not hold in MCDA as usually there is no option that dominates all the others with respect to all criteria considered. Hence, the goal of MCDA is not confined to finding an optimal solution but rather help the DM bring more transparency into the problem and guide him/her towards a satisfactory compromise solution, thus being a more realistic approach to decision making.

The history of MCDA dates back to the 18th century, when the French mathematician and

philosopher Marquis de Condorcet published his "Essay on the Application of Analysis to the Probability of Majority Decisions". However, it took almost two centuries for MCDA to emerge in its own right as a research area in decision analysis. The theory of revealed preferences proposed by Samuelson (1938), the birth of game theory (Von Neumann and Morgenstern, 1944), Arrow's social choice theory and his impossibility theorem (Arrow, 1951), the mathematical psychology analysis of individual decisions by Luce and Raiffa (1957), and Simon's bounded rationality theory (Simon, 1972), created a fertile ground for MCDA to lay its roots. The theory of bounded rationality was perhaps the one which most influenced this research area, since it introduced the idea that the rationality of decision makers is limited by the resources they have available, be it time, information, or even cognitive ability. As a consequence, decision makers usually choose a satisfactory solution rather than an optimal one. Related ideas were further reinforced by the work of Amos Tversky and Daniel Kahneman on judgements and decision making. Tversky and Kahneman (1974) recognised that human judgements can be limited, distorted, and prone to bias, especially when faced with uncertainty and problems that require the processing and analysis of large amounts of complex information. A well known example of cognitive bias in decision making is *anchoring*, a notion introduced by Tversky and Kahneman (1974) that describes the human tendency to rely too heavily on the starting information (the anchor) when making decisions. This anchor further influences the subsequent judgements as the information received *a posteriori* is interpreted in relation to the anchor, thus producing biased decisions. A few years prior to the publication of some of these theories and findings, namely in the 60's, researchers started to devote themselves to the development of MCDA methods and techniques in an attempt to overcome the limitations posed by human judgement when facing problems with multiple criteria. Due to its relevance, MCDA quickly evolved and established itself as an active research area in the 70's, and a large number of multicriteria methods have been developed in the 80's and 90's. These MCDA approaches evolved around two schools of thought: the European Multicriteria Decision Aid (MCDA) and the American Multicriteria Decision Making (MCDM). In this document, the term MCDA will be used regardless of the school involved. The former school devises relational or outranking methods based on the constructive approach, whose goals are to build rationality models tailored to the decision maker and give recommendations according to these models. The latter school develops functional methods based on the normative approach and, thus, tries to approach an ideal solution based on a rationality model derived from a set of axioms, or norms of rational behaviour. Regardless of the school of thought, the majority of MCDA methods proposed so far share a number of characteristics. First, these methods aim at providing guidance for carrying a structured, transparent, and efficient decision-making process that covers all the important factors that are likely to differentiate alternative courses of action. Secondly, the application of MCDA in real-world problems helps increase the confidence of the decision makers in their decisions, by helping them reach a solution that complies with their preferences and system of values. Thirdly, due to the

interactive and iterative nature of the MCDA process, its application in real-world scenarios may prove to be a daunting and time-consuming task, which requires a significant endeavour from both analysts and decision makers. Therefore, MCDA is more suitable for supporting decision making regarding problems of high complexity and that may possibly lead to long term impacts (Brito et al., 2010). In this chapter, we adopt the definitions of *decision makers* and *analysts* proposed by Belton and Stewart (2002), thus regarding the *decision maker* as the one who has the responsibility for the decision, and the *analysts* (also called facilitator or researcher) as those who guide and aid the decision makers in the process of reaching a satisfactory decision.

MCDA methods assist the DM in the process of identifying the most desirable option(s), from a set of possible options (explicitly or implicitly defined), when there are multiple and often conflicting objectives (for instance, maximize quality and minimize costs), which are potentially incomparable and incommensurable. Objectives are measured in terms of different evaluation criteria. A criterion is a tool for evaluating and comparing options under a well defined point of view. Options (also referred to as alternatives, actions, scenarios, plans, and programs) are defined as the object of the decision. They distinguish themselves by the extent to which they achieve the objectives, since usually none of the options has the best performance for all criteria. Depending on the typology of the MCDA problem at hand, these options can be implicitly found by solving a mathematical model (Multi-Objective Decision Making) or they can instead be explicitly known (Multi-Attribute Decision Making). In the former, the decision space is continuous (*e.g.*, Goal Programming and Multi-Objective Programming), whereas in the latter the decision space is discrete and the number of options is finite and predetermined (*e.g.*, ELECTRE, PROMETHEE, AHP, MAUT, MACBETH). Note that these options do not need to be implementable *a priori*. Evaluation criteria (also referred to as *attributes*) are performance measures (qualitative or quantitative) that are ranked by the DM in terms of his/her perceived importance and which are considered together when appraising the options. By explicitly assessing the performance of different options, based on the integration of objective measurement with subjective value judgement, MCDA techniques unavoidably lead to more informed and transparent decisions. The goal of MCDA is not to prescribe the "best" decision to be chosen, but to help decision makers select an option, or a short-list of good options, that best fit their needs and is coherent with their preferences and general understanding of the problem (Stewart, 1992). This option corresponds to the best compromise solution rather than to an optimal solution. This is somehow related to the significant degree of subjectivity inherent to any decision process, and the need to incorporate this subjectivity in the assessment of the possible solutions (Lu et al., 2007).

In current days, most of the decisions are taken by a Group of Decision Makers (GDM), rather than by a single individual. As stated by Zionts (1979), the idea of an omniscient decision maker is often a myth, because very seldom the DM knows the right course of action before getting acquainted with key information provided by other people. However, solving decision problems

that involve multiple decision makers poses additional problems to the already challenging MCDA research area. The difficulty of group decision making arises from the fact that the GDM is comprised of individuals with different value systems and distinct perceptions of how a problem should be tackled and solved. Several procedures were proposed to aggregate the individual judgements and points of view of a set of DMs, regarding the attractiveness of potential options, into a collective value. Forman and Peniwati (1998) argued that the two most useful procedures to address this problem are the following: the first procedure, known as the *aggregation of individual judgements*, assumes that the group acts together as a unit, thus engaging in discussion until a consensus is achieved; on the other hand, the second procedure, referred to as *aggregation of individual priorities*, assumes the group acts as separate individuals, each one expressing their own preferences. In the latter, the collective preference can be identified using vote counting schemes (for instance, plurality method, Hare voting, Coombs voting), ranking systems and the method of analysis of individual priorities, just to name a few.

The views of academics, such as Belton and Stewart (2002) and Seydel (2006), agree that MCDA prompts learning and a better understanding of the perspectives of the DMs themselves and the perspectives of the remaining key players involved in the decision process. Learning and understanding of the problem is mostly achieved by stimulating reflection, sharing of ideas, and discussion about the problem at hand. These activities increase the transparency of the decision-making process and might hasten the reaching of consensus. Thus, MCDA can act as a methodology to document, support, and justify decisions.

5.3 Methodology

The deployment of MCDA in real-world problems is an interactive and non-linear recursive process comprising several stages, which vary in number depending on the specific approach. Nevertheless, it is possible to outline the critical steps of a generic MCDA process that traverse the great majority of MCDA approaches.

The starting point of a MCDA process is both the establishment of a common understanding of the decision context and the identification and formulation of the decision problem. This is a crucial step, which involves the DM and other key players that are able to make significant contributions to the MCDA process by sharing their expertise. The second and third steps of the process comprise the identification of the options, as well as of the decision criteria that are relevant for appraising them. The step that follows is the assignment of relative importance weights to the chosen criteria. These weights can be determined directly (rating, ranking, swing, trade-off) or indirectly (centrality, regression, and interactive). Afterwards, the expected performance of each option against the criteria is assessed. This step involves a scoring/preference elicitation procedure to establish preferences between options, which varies according to the adopted MCDA

method. The most common scoring/elicitation procedures are: trade-offs, lotteries, direct rating, and pairwise comparisons. This step is typically supported by an evaluation matrix (also referred to as consequence matrix, options matrix, performance matrix or simply multicriteria table) that sets out the performance of each option against a common set of criteria. The global scores of the options are then obtained by assessing, for each criterion, the values associated with their consequences, leading to a set of multicriteria scores, one for each option. The set of options is usually ranked based on these global scores, from best to worst. Eventually, the process may also involve a sensitivity analysis of the results to changes in scores or criteria, in order to infer the stability and robustness of the outcome of the MCDA process. Finally, the evaluation and trade-offs involved in the options considered are provided and discussed with the DM. It is important to note that the results yielded by a MCDA process are not prone to generalisations, in the sense that they only apply to the set of options that were evaluated (Dooley et al., 2009), and are hard to verify or repeat since they are highly dependent on the specific process and the interactions among the actors involved (Hobbs and Meier, 2012).

Although several MCDA methods have been proposed over the years, this chapter describes the AHP and the PROMETHEE, as well as the AHP-PROMETHEE approaches, since these are the ones used in this study.

5.3.1 AHP

One of the most prevalent and popular approaches for MCDA is the AHP. This problem-solving framework was originally developed by the mathematician Thomas Saaty (1977), in the late 70's, and belongs to the family of normative methods of the American school of thought. Albeit the severe criticism and heated debate that AHP has been subjected to by MCDA academics (see, for instance, Belton and Gear (1983), Schoner and Wedley (1989), Dyer (1990), and Bana e Costa and Vansnick (2008)), its widespread application reflects its generalised acceptance by both the academic community and the practitioners community.

The basic idea behind the AHP is converting subjective assessments of relative importance into a set of overall scores and weights. The assessments are subjective since they reflect the perception and preferences of the DM, and relative because they are based on pairwise comparisons of criteria or options. The AHP converts these subjective assessments of the DM into ratio-scale weights that are further combined into linear additive weights for the associated options (Forman and Gass, 2001). There are two goals here: derive and estimate the priorities, or weights, of criteria and establish the relative performance scores for the options in each criterion.

According to Forman and Gass (2001), the AHP is a methodology based on three main functions: (i) structuring complexity, (ii) measurement on a ratio scale, and (iii) synthesis. The first function is related to the decomposition of the decision problem into a hierarchy of sub-problems, by arranging the relevant factors of the problem into a hierarchic structure that descends from a global goal

Table 5.1: Nine-point intensity scale proposed by Saaty (1986) within the AHP method, as given by Turcksin et al. (2011).

Definition	Intensity of Importance
Equal importance	1
Moderate importance	3
Higher importance	5
Much higher importance	7
Complete dominance	9
Intermediate values	2, 4, 6, 8
Reciprocals	$1/2$, $1/3$,..., $1/9$

to criteria, sub-criteria, sub-sub-criteria, in successive levels, until it reaches the last level, which corresponds to the options. The second function comprises the elicitation of pairwise comparison judgements from the DM. This is carried out by asking the DM to compare pairs of criteria and to indicate, for each pairwise comparison, their relative contribution to the immediately higher-level element in the hierarchy. These relative evaluations are scaled from one to nine using the nine-point intensity scale proposed by Saaty (1977) (please see Table 5.1) and arranged into a squared positive reciprocal $n \times n$ matrix, with n being the number of criteria. The vector of priorities, or criteria weights, is further derived by applying Saaty's eigenvector method to the comparison matrix. This vector of priorities is the normalised principal right eigenvector of the pairwise comparison matrix corresponding to the maximum eigenvalue of the matrix, and is expressed on a ratio scale.

The use of pairwise comparisons to input qualitative information is a major advantage since these are straightforward, intuitive, and a convenient means to extract subjective information from the DM. In fact, the idea of using pairwise comparisons is the reason behind the method's popularity. However, these pairwise comparisons may lead to inconsistencies in the judgements. Prior to acceptance, a consistency verification step is carried out to evaluate the degree of consistency of the DM's judgements, which were given in the form of pairwise comparisons. This is achieved by computing a Consistency Ratio (CR) that measures the consistency of the DM's judgements with respect to a large sample of random judgements (Saaty, 1977, 1986, 1990). CR is computed as follows:

$$CR = \frac{CI}{RCI}, \quad \text{with} \quad CI = \frac{\lambda_{max} - n}{n-1}, \tag{5.1}$$

where λ_{max} denotes the maximum eigenvalue of the pairwise comparison matrix, n represents the squared matrix size (in this case, the number of criteria), and the acronyms CR, CI, and RCI stand for Consistency Ratio, Consistency Index, and Random Consistency Index, respectively. RCI is obtained from a randomly generated pairwise comparison matrix of size n (please see Table

5.2). According to Saaty (1977), if $CR > 0.10$ the DM's input reveals an unacceptable degree of inconsistency (non-transitivity) and, thus, his/her judgements are deemed unreliable. In such cases, the entries of the pairwise comparison matrix need to be amended before proceeding to the next step of the AHP.

Table 5.2: Random consistency index for the most common matrix sizes (Saaty, 1986).

Size of the matrix (n)	3	4	5	6	7	8
Random Consistency (RCI)	0.58	0.90	1.12	1.24	1.32	1.41

The third step of the AHP is the evaluation of the options, or synthesis. This can be performed using two approaches: one based on the relative measurement of options, and another based on absolute measurements of these options. In the first approach, separate pairwise comparisons are conducted on the set of options in each criterion (and sub-criterion, sub-sub-criterion, *etc.*, if applicable) in order to elicit their performance scores. In the latter approach, the options are simply rated for each criterion, by identifying the grade that best describes them. Afterwards, an aggregation step yields the final results of the AHP, which are the orderings of the options based on a global indicator of priority. The aggregation step summarises the performance of the options over the criteria into a set of priority scores (one for each option). The option with the largest priority score is the most desirable compromise solution.

The wide applicability of the AHP is partly justified by the following distinctive advantages of the method: (i) modelling and structuring of complex decision problems and (ii) determination of criteria weights. The former is important because it allows decomposing a complex problem into sub-problems by means of a hierarchical decision tree. This kind of representation makes it easier for the DM to obtain a clearer view of the complex relationships inherent to the decision problem, contributing to a better understanding of the decision process. The latter is also an advantage since the criteria weights can be determined by the eigenvector method, which returns weights with a higher level of coherence, consistency, correlation, and accuracy, than weights determined on the basis of intuition or domain knowledge (Bogdanovic et al., 2012). However, despite these advantages, AHP displays several drawbacks when applied to real decision problems (Belton and Gear, 1983; Schoner and Wedley, 1989; Dyer, 1990; Goodwin and Wright, 2004; Macharis et al., 2004), such as: (i) the compensatory effect of trade-offs and (ii) the time involved in the process of gathering the DM's evaluations. The first drawback means that the degradation of performance on a particular criterion is compensated by good performance in others as a consequence of the complete aggregation of the additive type embedded in the method. As a consequence, detailed information can be lost in the aggregation process (Macharis et al., 2004). This compensatory effect is amplified in the presence of a large number of criteria. The second drawback is related to the substantial amount of pairwise comparisons that need to be performed by the DM when

dealing with a large number of criteria (over ten) and/or options. This requirement of the method is detrimental since it can turn the preference elicitation step into a cumbersome and time-consuming task. Criticism has also been made regarding the potential inconsistency of the nine-point intensity scale (Murphy, 1993), the rank reversal problem (Belton and Gear, 1983; Barzilai et al., 1987), and the inappropriateness of the CR for detecting the existence (or the non-existence) of a numerical scale satisfying the Condition of Order Preservation (COP) (Bana e Costa and Vansnick, 2008).

For these reasons, in this study the AHP was complemented with another MCDA method, the PROMETHEE, so that a more confident evaluation and analysis can be provided to the DM. The advantages of combining both approaches will be explained in Section 5.3.3.

5.3.2 PROMETHEE

The PROMETHEE is another widespread family of methods for addressing multicriteria decision problems (see, for instance, Behzadian et al. (2010) for a comprehensive review of its applications). It is well known for being a quite simple ranking method to design and apply (Gervásio and Simões da Silva, 2012). The PROMETHEE family belongs to the European, or French, school of thought, which relies on the concept of partial aggregation, as opposed to the complete aggregation proposed by the American school. Methods of partial aggregation are better known as outranking methods. The theory behind these methods was first introduced by the French Professor Bernard Roy (1968).

The PROMETHEE requires both criteria and options to be specified beforehand and is based on pairwise comparisons of options along each criterion, *i.e.*, the DM assigns scores to measure the performance of the options comparatively to each criterion. These data are summarised and arranged into an evaluation matrix, which is regarded as the method's starting point. The following stages encompass (i) the weighting given to each criterion and (ii) the definition of the preference function used by the DM when appraising the contribution of the options in terms of each criterion.

Several methods have been proposed to help determine the weights of the criteria (an overview of the classic methods is provided by Eckenrode (1965)). In PROMETHEE, the weighing step is a non-trivial task that highly depends on the DM's ability to properly weigh the criteria, since no formal guidelines for determining these weights are provided. Therefore, in an attempt to reduce the subjectivity associated with this stage, in this work the AHP method is used to obtain the weights.

The PROMETHEE also requires the specification of a preference function for each criterion g_j ($j = 1, ..., n$, where n denotes the total number of criteria). A preference function $P_j(a,b)$ translates the deviation between the appraisals of two options (a and b, $\{a,b\} \in A$, with A representing a finite set of m possible options), on a particular criterion g_j, into a preference degree ranging from 0 to 1. This preference degree is a non-decreasing function F_j of the observed deviation $d_j(a,b) = g_j(a) - g_j(b)$, as shown in the following equation:

$$\forall_{a,b\in A}, \qquad P_j(a,b) = F_j\left[g_j(a) - g_j(b)\right], \qquad 0 \leq P_j(a,b) \leq 1. \tag{5.2}$$

Thus, smaller deviations will contribute to weaker degrees of preference, whereas larger deviations will contribute to stronger ones. Note that for criteria to be minimized, the preference function should be reversed (Equation (5.3)).

$$\forall_{a,b\in A}, \qquad P_j(a,b) = F_j\left[-d_j(a,b)\right], \qquad 0 \leq P_j(a,b) \leq 1. \tag{5.3}$$

Brans and Vincke (1985) provide some guidance for this stage by proposing six basic configurations of preference functions, namely: linear, usual, level, U-shape, V-shape, and Gaussian functions. The first five functions rely on either the indifference threshold **q**, the preference threshold **p** or both, while the Gaussian function depends on a single parameter **s**.

Indifference threshold q**:** the largest deviation considered negligible by the DM, on a particular criterion. Within the indifference threshold, the DM does not perceive any difference between two options. This threshold is a small value with respect to the scale of measurement of the criterion.

Preference threshold p**:** the smallest deviation considered as sufficient to generate a full preference over one of the options. This threshold is a large value with respect to the scale of measurement of the criterion. Note that **q** has to be smaller than **p**.

Gaussian threshold s**:** an intermediate value between **q** and **p** that represents the inflection point of the Gaussian function.

The next step of the PROMETHEE uses the previous information, namely the criteria weights w_j and the preference functions $P_j(a,b)$, to compute an overall preference index $\pi(a,b)$ (Equation (5.4)) that expresses the intensity of preference of option a over option b, when taking into account the whole set of criteria.

$$\forall_{a,b\in A}, \qquad \pi(a,b) = \sum_{j=1}^{n} w_j P_j(a,b), \qquad \sum_{j=1}^{n} w_j = 1. \tag{5.4}$$

This preference index serves as a basis to compute "core" quantities, namely the outranking flows. The positive (or leaving) outranking flow $\phi^+(\mathbf{a})$ (Equation (5.5)) measures the degree to which a given option a outranks all the other options. An option a outranks another option b if a outperforms b on enough criteria of sufficient importance and, at the same time, does not exhibit a much inferior performance on any other criterion. Likewise, the negative (or entering) outranking flow $\phi^-(\mathbf{a})$ (Equation (5.6)) expresses how much a given option a is dominated (or outranked) by all the other options. The higher/smaller the positive/negative flow, the better the option. The

balance between these flows is represented by the net outranking flow $\phi(\mathbf{a})$ (Equation (5.7)), which is a dimensionless quantity. A higher value of $\phi(\mathbf{a})$ reflects a higher attractiveness of option a. The higher the positive flow $\phi^+(\mathbf{a})$, *i.e.*, the closer to 1, and the smaller the negative flow $\phi^-(\mathbf{a})$, *i.e.*, the closer to zero, the better the option.

$$\phi^+(a) = \frac{1}{n-1}\sum_{x\in A}\pi(a,x). \tag{5.5}$$

$$\phi^-(a) = \frac{1}{n-1}\sum_{x\in A}\pi(x,a). \tag{5.6}$$

$$\phi(a) = \phi^+(a) - \phi^-(a). \tag{5.7}$$

Although the PROMETHEE family encompasses several methods, the main ones are (i) PROMETHEE I partial ranking, (ii) PROMETHEE II complete ranking, and (iii) GAIA. PROMETHEE I partial ranking generates a ranking of options that, in some cases, is incomplete (or partial) since it allows indifference and incomparability situations between options. This way, the PROMETHEE I is able to avoid trade-offs between scores on criteria, which is likely to happen in, for instance, the AHP. The ranking generated by the PROMETHEE I is based on the simultaneous comparison of positive and negative outranking flows ($\phi^+(\mathbf{a})$ and $\phi^-(\mathbf{a})$), for each pair of options (a and b, $\{a,b\} \in A$). Thus, option a is preferred to option b (denoted as aPb) if a concomitantly obtains a higher (or equal) positive flow and a lower (or equal) negative flow than b, as shown in Equation (5.8). If all the considered options satisfy these preference conditions, the PROMETHEE I returns a complete ranking. Yet, sometimes only a partial ranking can be generated because there are pairs of options that cannot be compared or that are indifferent. Incomparability (denoted as aRb) occurs when the comparison of two options a and b is difficult, typically because they exhibit quite distinct profiles (option a gets high scores on a subset of criteria in which b is weak, and vice versa). This can be translated into the set of conditions presented in Equation (5.8). In turn, two options are said to be indifferent (denoted as aIb), if they both obtain equal scores on $\phi^+(\mathbf{a})$ and $\phi^-(\mathbf{a})$, as presented at the end of Equation (5.8).

$$\begin{cases} aPb \quad \text{if} \begin{cases} \phi^+(a) > \phi^+(b) \text{ and } \phi^-(a) < \phi^-(b), \text{or;} \\ \phi^+(a) > \phi^+(b) \text{ and } \phi^-(a) = \phi^-(b), \text{or;} \\ \phi^+(a) = \phi^+(b) \text{ and } \phi^-(a) < \phi^-(b). \end{cases} \\ aRb \quad \text{if} \begin{cases} \phi^+(a) > \phi^+(b) \text{ and } \phi^-(a) > \phi^-(b), \text{or;} \\ \phi^+(a) < \phi^+(b) \text{ and } \phi^-(a) < \phi^-(b). \end{cases} \\ aIb \quad \text{if} \quad \phi^+(a) = \phi^+(b) \text{ and } \phi^-(a) = \phi^-(b). \end{cases} \tag{5.8}$$

Conversely, PROMETHEE II relies directly on the net preference flow $\phi(\mathbf{a})$ to rank the options. Hence, it always provides a complete ranking of the options and, thus, all options are deemed comparable.

One benefit of the PROMETHEE is that the output can be graphically represented using a technique called GAIA. The GAIA plane is a geometrical representation of the relative position of the options in terms of the contribution to the various criteria. This representation directly results from applying the Principal Component Analysis to the variance-covariance matrix associated with the matrix containing all the unicriterion net flows. These flows are computed according to Equation (5.9). The n-dimensional criteria space is projected onto a two-dimensional space yielded by the two first principal components (linear combinations of the original criteria) so as to preserve as faithfully as possible the original multidimensional information.

$$\phi_j(a) = \frac{1}{m-1} \sum_{a \neq b} \left[P_j(a,b) - P_j(b,a) \right]. \tag{5.9}$$

Similarly to the AHP method, PROMETHEE is not a flawless method. The most commonly reported weaknesses of PROMETHEE are: (i) the absence of guiding principles for structuring the decision problem and (ii) the lack of specific guidelines for eliciting criteria weights (Macharis et al., 2004). The former prevents a lucid understanding of the problem by the DM, which proves particularly important when dealing with decision problems that involve several criteria levels (*e.g.*, criteria, sub-criteria, sub-sub-criteria) and a large number of options. The latter drawback can be problematic since it relies on the assumption that the DM is able to judiciously weigh the criteria. However, this ability is not always guaranteed in practice because criteria weighing is a challenging task and highly affected by the skillfulness of the decision maker.

The aforementioned disadvantages of PROMETHEE correspond to the advantages of the AHP method. Thus, these drawbacks of PROMETHEE can be obviated by integrating it with the AHP method, as will be explained next.

5.3.3 The Hybrid Approach

As described at the beginning of this section, the implementation of a MCDA methodology comprises several stages. Although it has been common practice to choose a single MCDA method (*e.g.*, AHP, ELECTRE, PROMETHEE) to carry out all the steps of the MCDA methodology, it is also possible, and becoming popular, to harness the strengths of different methods and use each one of them to perform the steps they are especially good at. By doing so, it is expected to create a more powerful approach, able to produce results that reflect more closely the DM's implicit preferences and ultimately improve the reliability and robustness of the MCDA process. This is the idea behind the combination of MCDA methods of different nature: the AHP, which is a functional method based on the normative approach, and the PROMETHEE, which is a relational method based on

the constructive approach. The former was applied to the MCDA stage related to the structuring and modelling of the decision problem, whereas the latter was used in the MCDA stage concerned with the elicitation and aggregation of preferences.

The idea of strengthening PROMETHEE with useful features of the AHP was first proposed by Macharis et al. (2004). The authors suggested combining the abilities of the AHP to both structure complex decision problems and determine criteria weights, so as to overcome the main weaknesses of the PROMETHEE. This idea attracted the attention of several researchers, and a number of case studies using the AHP-PROMETHEE hybrid approach were reported in the literature. The main differences between these case studies and the one presented here is the field of application and the targeted problem. The AHP-PROMETHEE has been used to solve certain problems such as information systems outsourcing decisions (Wang and Yang, 2007), equipment selection (Dağdeviren, 2008), multi-instrumentality policy package selection (Turcksin et al., 2011), mining method selection (Bogdanovic et al., 2012), supply chain risk prioritisation (Venkatesan and Kumanan, 2012), comparative assessment of infrastructures from a sustainability viewpoint (Gervásio and Simões da Silva, 2012), crop planning under different economic conditions (Agha et al., 2012), task prioritisation in disassembly workstations (Avikal et al., 2013), determination of the most suitable areas for flood spreading and artificial recharge of aquifers (Nasiri et al., 2013), selection of an efficient 3PL firm for a manufacturer (Bansal and Kumar, 2013), and selection of municipal solid waste treatments (Herva and Roca, 2013). Nevertheless, to the best of our knowledge, to date no work has used the AHP-PROMETHEE to solve problems arising in the paint shop of an automobile assembly plant. Variations of the originally proposed AHP-PROMETHEE were also developed so as to accommodate the specificities of certain decision problems, see *e.g.*, Kara (2011), Taha and Rostam (2012), Ilangkumaran et al. (2013), Avikal et al. (2014), and Roodposhti et al. (2014).

As previously mentioned, our methodological approach hybridises the AHP and the PROMETHEE in order to obviate, or even overcome, some faults of each one of them. In Section 5.3.1 we pointed out two drawbacks of the AHP, namely, (i) its compensatory effect and (ii) the time involved in the process of gathering the DM's evaluations. By joining the two methods, the first issue of the AHP is resolved by the non-compensatory property of the PROMETHEE I (Perny, 1998), which is a consequence of its outranking nature. Contrary to what happens in the AHP, in PROMETHEE I the degradation of performance in a given criterion cannot be compensated by good performance in others. Instead, the performance of the options on each criterion (whether good or bad) is preserved, thus preventing information loss. The hybridisation also eliminates the second concern of the AHP. While the AHP requires the DM to perform a substantial amount of pairwise comparisons, which can become very large when considering many options and/or criteria, the PROMETHEE significantly reduces this number due to its ability to indirectly achieve a synthesis. In addition to these advantages, the PROMETHEE improves the AHP by associating a preference value based on

attractiveness functions to each criterion that reflects more closely the DM's preference structure. On the other hand, the integration of both methods mitigates two major disadvantages associated with the PROMETHEE, namely (i) its lack of problem structuring and (ii) the ambiguous procedure for weighing criteria (*cf.* Section 5.3.2). These happen to be two of the most consensual strengths of the AHP. The first drawback can be resolved by incorporating an AHP-like hierarchy tree structure in the PROMETHEE. The second problem can be obviated by determining criteria weights using the eigenvector method of the AHP. Such procedure yields more objective criteria weights (Bogdanovic et al., 2012).

Note however that both methods suffer from the rank reversal problem (Belton and Gear, 1983; Barzilai et al., 1987). This means that, in some cases, the ranking of the options can be reversed when a new option is added or removed from the decision model after preferences have been given (Macharis et al., 2004). This problem can be mitigated by devoting special attention to the problem structuring phase in order to ensure that both the options and criteria considered in the decision model are correct.

The adopted AHP-PROMETHEE methodology is comprised of two broad stages, which are further partitioned into detailed steps: (i) AHP problem modelling and elicitation of criteria weights and (ii) PROMETHEE computations.
The first stage encompasses four steps:

Step 1: Identify the painting plans (*i.e.*, the options);

Step 2: Select an exhaustive and consistent family of criteria that are relevant to the decision context;

Step 3: Construct the AHP decision tree;

Step 4: Determine the criteria weights using the elicitation procedure and eigenvector method of the AHP.

In the second stage, the PROMETHEE is used to obtain global priorities of the options. This stage comprises the following five steps:

Step 5: Compute the positive and negative outranking flows for each option by following this stepwise procedure: build the evaluation matrix; determine the preference functions and the corresponding parameters for each criterion; compute the deviations for all pairs of options a and b for each criterion; apply the selected preference functions to these deviations in order to quantify the preference among options; compute the overall preference index as a weighted sum of preference functions, over all criteria.

Step 6: Generate the PROMETHEE I partial ranking using the positive and negative outranking flows;

Step 7: Compute the net flows and generate the PROMETHEE II complete ranking;

Step 8: Visually analyse the results using the GAIA plane;

Step 9: Perform a sensitivity analysis by determining and analysing stability intervals for criteria.

After the successful completion of these steps, the results are discussed with the DM so he/she can proceed with the decision.

5.4 Case Study: Evaluating the Vehicle Painting Plans

The target of this case study is one of Toyota's assembly plants whose main function is to perform the welding, painting, and final assembly of a specific automobile model and corresponding family of products (single cabin and double cabin vehicles). These functions correspond to the three main stages of an assembly process, namely the body shop, the paint shop, and the trim and assembly shop (Ulgen and Gunal, 1998). This plant experienced a decline in its demand and a subsequent sales drop as a consequence of the pernicious effects of the automobile industry crisis of 2008-2010. As a result, the plant is producing below capacity. However, this scenario may soon change since the plant applied for the production of a new automobile model, whose forecasted production would require the plant to operate at full capacity. Thus, there is the need to optimise the processes. Since the painting activity is (i) one of the most important and complex activities in the automobile manufacturing (Geffen and Rothenberg, 2000; Li et al., 2007), (ii) a bottleneck in this specific plant, and (iii) responsible for the highest costs (according to the information provided by the operations manager, approximately 70% of the total expenditures of the entire assembly plant stem from the paint shop and related operations), the management considered this sector to be the most critical to conduct a MCDA.

It follows an overall description of the assembly line of the plant, with a special focus on the painting operations.

The vehicle components are delivered to the plant in batches including the necessary components for five vehicles. After assembling, the welded body is directed to the paint shop. The paint shop comprises a production line that is made up of 17 work stations (please see Figure 5.1). When the vehicles' body (or simply *cabins* in this case) are transferred to the paint shop, they are first submitted to a prewash. The main process begins at the next station, where the surface of the cabins is cleaned and prepared for the subsequent application of organic coatings through a chemical pretreatment. Then, the surfaces of the vehicles' body are washed again and further submitted to electrocoating. Afterwards, these are dried in an oven, which bakes the coat of paint, and subjected to a manual inspection. If a defect is found, it is corrected by manual sanding. Next, sealing and PVC (stands for *Polyvinyl Chloride*) are applied to prevent humidity and corrosion. The sealing is dried in another oven, and then the cabins are wiped to be subsequently subjected to a primer

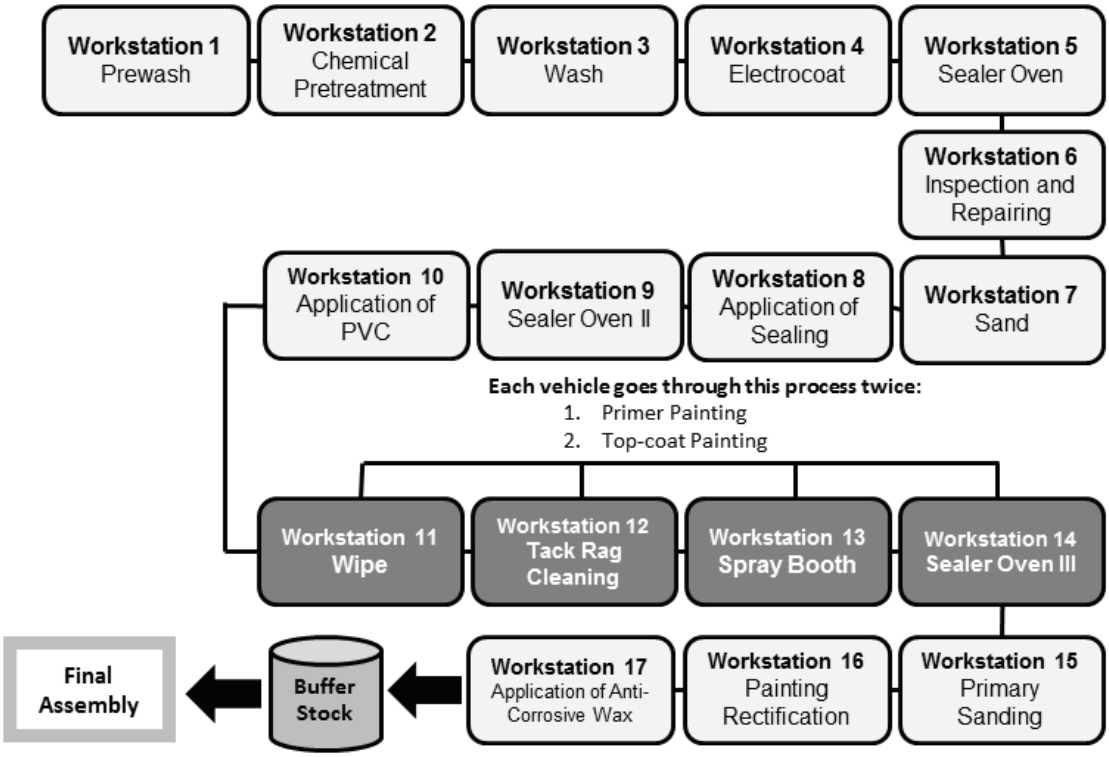


Figure 5.1: Illustration of the complete job flow in the automobile paint shop.

painting, in a spray booth, and dried in an oven. The operations performed at workstations 11, 12, 13, and 14, in Figure 5.1, are repeated when applying the top-coat. The process continues with the manual inspection of the physical aspect of the painted surface. As before, if defects are detected, they are corrected by manual sanding and rectification. The painting activities end with the application of anti-corrosive wax. The painted cabins are then stocked in a buffer stock until they are forwarded for final assembly.

In the plant under study, the only way to improve the operational efficiency of the painting activities is by optimising the vehicle painting plans, *i.e.*, the schedule. The daily production of the plant is intended to meet the clients' demand. The client base is international, which reflects on a diversity of vehicle specifications (*e.g.*, vehicle colour preference varies according to the client's country). The plant is a lean manufacturing system with high throughput and very little inventory, and relies on a built-to-order production system. Therefore, the production is carried out according to the clients' specifications which typically correspond to different combinations of colours (*e.g.*, white, red, blue, gray) and cabin versions (single or double). The analysis of these combinations allowed the management to define a set of the most frequently used painting plans to schedule the paint shop activities. This set corresponds to the options of the decision problem, which will be detailed next.

Due to its technical nature, the painting activity involves a set of sequential operational procedures whose overall performance is influenced by the colour planning of the vehicle cabins. These colour plannings correspond to what we call *painting plans*. These painting plans are defined as a combination of the type of cabin of the automobile model painted in a given day (single cabin or mixed cabin), with the number of distinct colours needed to paint these vehicle cabins. The expression *single cabin* indicates that the units painted in a given day are all single cabins, whereas the expression *mixed cabin* refers to daily painting plans including more than one cabin version (*i.e.*, both single and double cabins are painted in a given day). Note that each vehicle cabin is painted using a single colour. However, during a given day, the paint shop might use different colours to paint different vehicle cabins, according to the clients' specifications. Thus, a painting plan also considers the number of distinct colours used in the paint shop in a given day. After going through the paint shop stage, the painted cabins are stored in a buffer stock before being forwarded for final assembly. It is due to the existence of this buffer stock, that is possible to adjust the painting plans schedule in order to achieve a better overall performance of the whole assembly process (*e.g.*, more quality, less energy consumption).

Enhancing the performance of the painting activities requires the evaluation of several painting plans using multiple criteria so as to capture the complexity of the process. The evaluation involves the systematic structuring of the different aspects of the painting plans, the examination of these options, and the assessment of their ability to jointly achieve a set of objectives set by the DM. These objectives are typically measured in terms of possibly conflicting criteria. It is thus necessary

to compromise among these criteria to find a ranking of the set of painting plans that complies with the DM's objectives and preferences, and can be used to improve the efficiency of the paint shop.

The main goals of this case study are the following: use a hybrid MCDA approach to (i) systematically structure the information regarding the decision problem, (ii) assist the DM in the process of assessing the most frequently used painting plans in terms of the evaluation criteria, (iii) sort the painting plans in decreasing order of preference, and (iv) identify the painting plan which contributes the most to the paint shop optimisation.

Note that, since a MCDA approach is used, the decision making lies in the hands of the DM not in the hands of the analysts. Therefore, the procedure and output of the AHP-PROMETHEE method only serve the purpose of guiding the DM through the decision-making process, using as a basis a combination of factual data and subjective judgements.

5.4.1 AHP Problem Modelling and Elicitation of Criteria Weights

Applying the MCDA to this decision problem involved the operations manager of the assembly plant and the paint shop team (henceforth Decision Maker, or simply DM). Although there are several people involved in the decision-making process, they act as if they were a single decision maker, since the answers provided represent the consensual views and preferences of both the manager and the paint shop team. Hereafter, we will refer to the decision maker as a single entity. A number of face-to-face meetings with the DM were convened so as to understand the decision context and to gather information regarding the decision problem, the possible set of options, and the relevant set of criteria.

As previously mentioned, the major goal of the DM is to optimise the global scheduling of the paint shop of the assembly plant by optimising the use of the vehicle painting plans. The starting point of this case study already included the set of options, *i.e.*, the most frequently used painting plans that were previously defined by the DM. In the first meeting, the major work done with the DM was to clarify the aims to be considered for the evaluation of the plans, *i.e.*, to improve the paint shop efficiency and decrease costs, while ensuring desirable quality levels. Once this has been accomplished, the DM had to evaluate the aforementioned plans, for which a set of criteria had to be specified. They were able to present their objectives, - decrease the number of painting defects while keeping the paint consumption at low levels, reduce the energy consumption, and increase the number of painted vehicles in a day -, and then together we defined how to measure them. That is, we helped the DM to translate these broad objectives into specific criteria.

Afterwards, the DM supplied the necessary data to build the evaluation matrix (please see Table 5.3). This matrix provides objective and quantitative information regarding the performance of each option on each relevant criterion. The matrix entries represent the average values of criteria for each painting plan, which were computed based on a convenience sample of observations collected during the time period starting on June 5, 2012, and ending on November 21, 2013. Prior

Table 5.3: Evaluation matrix. The best values observed for each criterion are underlined. The analysis of these values reveals that none of the options achieves the best performance in all criteria, as expected.

Options/Criteria	**QI**	**EC**	**PC**	**NPV**
Unit	# Defects	kWh	Ink liters	# Vehicles
Max/Min	Min	Min	Min	Max
Weights AHP	0.6055	0.2296	0.1255	0.0394
PP-A (Single + 1 Colour)	<u>2.14</u>	<u>18</u>	1.59	10
PP-B (Single + 2 Colours)	2.19	23	<u>1.57</u>	14
PP-C (Single + 3 Colours)	2.58	21	1.64	14
PP-D (Single + 4 Colours)	2.33	23	1.60	<u>16</u>
PP-E (Mixed + 1 Colour)	3.0	21	1.81	13
PP-F (Mixed + 2 Colours)	2.52	21	1.92	14
PP-G (Mixed + 3 Colours)	3.45	25	1.92	15
PP-H (Mixed + 4 Colours)	3.20	25	1.88	<u>16</u>

to the computation of these average values, the data sample was preprocessed and statistically analysed. From an initial set of 112 daily observations, six observations were removed during the preprocessing stage, either due to the presence of many missing values or because they corresponded to anomalous working days in the assembly plant (*e.g.*, equipment malfunction). The statistical analysis was performed over the final sample, which comprised 106 daily observations. The detailed results of this analysis are provided in Appendix B. From the statistical analysis, we were able to conclude the following. First, the criteria show slightly asymmetric distributions, according to the analysis of Pearson's second skewness coefficient. Hence, it is acceptable to use the mean values, instead of the median values, for representing the data in the evaluation matrix. Additional experiments using the median values of the criteria corroborated this conclusion, since the PROMETHEE II complete ranking obtained for the average values and for the median values is very similar (only the two last ranking positions switched). Secondly, according to the analysis of Pearson's correlation coefficient, the criteria show weak or negligible correlation, thus ensuring that the data do not violate the assumption of independence of criteria. Thirdly, the criteria show acceptable levels of variability on each painting plan, based on the analysis of the Coefficient of Variation (CV). Among all criteria, the quality index is the criterion showing the highest relative variability, with its CV ranging from 0.29 (PP-D) to 0.52 (PP-E). Bearing in mind that real data seldom satisfy ideal theoretical properties or present themselves as we would like them to, a *CV* of 50% might be acceptable and even expected in real-life scenarios, such as the one we address. Here, a possible explanation for this *CV* value is the human component associated with the quality criterion. Knowing that a significant portion of defects arise as a consequence of the manual painting operations, which are performed by painters, the quality of the painting might be related

with the variability in the painters' performance. Based on the conclusions of the statistical analysis, we considered that the data available had acceptable quality to proceed with the empirical study. Note however that the results derived from this study cannot be generalised to other assembly plants, due to, for instance, the type of sampling used to collect the data (convenience sampling) and the statistical characteristics of the data.

The portfolio of options corresponds to a set of eight painting plans. For simplicity, in this chapter these will be referred to as PP-A (Painting Plan A), PP-B through to PP-H (**Step 1**, *cf.* Section 5.3.3). The final sample of 106 observations breaks down into the painting plans as follows: 23 observations for PP-A (21%), 22 observations for PP-B (21%), 13 observations for PP-C (12%), three observations for PP-D (3%), 19 observations for PP-E (18%), 16 observations for PP-F (15%), six observations for PP-G (6%), and four observations for PP-H (4%). These options can be concisely described as follows:

- **PP-A:** Daily plan characterised by the painting of *single cabins* using a *single colour*. This means that in this painting plan, in a given working day, all the vehicles are painted with the same colour (*e.g.*, gray) and are all of the same version (single cabins).
- **PP-B:** Daily plan characterised by the painting of *single cabins* using *two distinct colours*. This means that all the vehicle cabins painted in a given day are single cabins but some of them will be painted using another colour (*e.g.*, 60% are painted in red, while the remaining 40% are painted in white).
- **PP-C:** Daily plan characterised by the painting of *single cabins* using *three distinct colours*.
- **PP-D:** Daily plan characterised by the painting of *single cabins* using *four distinct colours*.
- **PP-E:** Daily plan characterised by the painting of *mixed cabins* using a *single colour*. This means that in this painting plan, in a given working day, all the vehicles are painted with the same colour (*e.g.*, blue) and are of different versions (single and double cabins are painted in the same day).
- **PP-F:** Daily plan characterised by the painting of *mixed cabins* using *two distinct colours*. This means that the vehicles painted in a given day are of different versions (single and double cabins) and are painted using two distinct colours.
- **PP-G:** Daily plan characterised by the painting of *mixed cabins* using *three distinct colours*.
- **PP-H:** Daily plan characterised by the painting of *mixed cabins* using *four distinct colours*.

These painting plans are appraised based on a set of criteria. These criteria are models that allow the development of preference relations between the options (*i.e.*, the painting plans). The

key criteria have been fully determined by the discussion between the analysts and the DM, and the choice was supported on the DM's objectives. Four quantitative criteria were considered after brainstorming sessions with the DM, namely: the Quality Index (QI), the Energy Consumption (EC), the Paint Consumption (PC), and the Number of Painted Vehicles (NPV) (**Step 2**). These criteria comprise the most relevant aspects for the assessment of the painting plans and are described as follows:

- **Quality Index (QI):** The quality index is given by the average number of defects per painted vehicle and, as the name implies, it is a proxy for the quality of the performed painting. Defects can arise as a result of the manual painting operations that are performed by painters, or as a consequence of the ink quality. This is a quantitative criterion based on qualitative data (types of defects). In this study, the quality index has been computed for each one of the identified painting plans by considering the total number of defects that were detected on the corresponding set of painted vehicles, during the time span under analysis. Thus, we obtain the defects as a ratio of the painted vehicles corresponding to each painting plan.

- **Energy Consumption (EC):** Energy consumption is a quantitative criterion, measured in kilowatts-hour (kWh) that indicates the gas energy directly consumed by the sealer ovens in a given day. This criterion has an impact on the costs of the paint shop, as well as on the environment.

- **Paint Consumption (PC):** This quantitative criterion reflects the direct cost of painting the vehicles (in terms of materials), and is given by the average ink litres used to paint a given vehicle. This is a weighted average whose weights reflect the proportion of ink consumption by colour, for each vehicle painting plan.

- **Number of Painted Vehicles (NPV):** The last criterion is the number of painted vehicles per day.

A summary description of the nature of these criteria is given at the top of Table 5.3.

Based on the information gathered, the decision problem was unbundled into its constituent parts using the AHP hierarchy tree structure comprising three levels (overall goal, criteria, and the different options or painting plans), as depicted in Figure 5.2 (**Step 3**). Despite the simplicity of the AHP tree for this specific decision problem, this visual representation helped the DM structure his thinking and more easily understand the decision problem and corresponding components.

After structuring the decision problem at hand, the DM was asked to assess the relative importance of the identified criteria based on pairwise comparisons, so as to obtain the criteria weights. These weights are non-negative numbers, are independent of the measurement units of the criteria, and are determined such that higher weights reflect higher importance. These weights

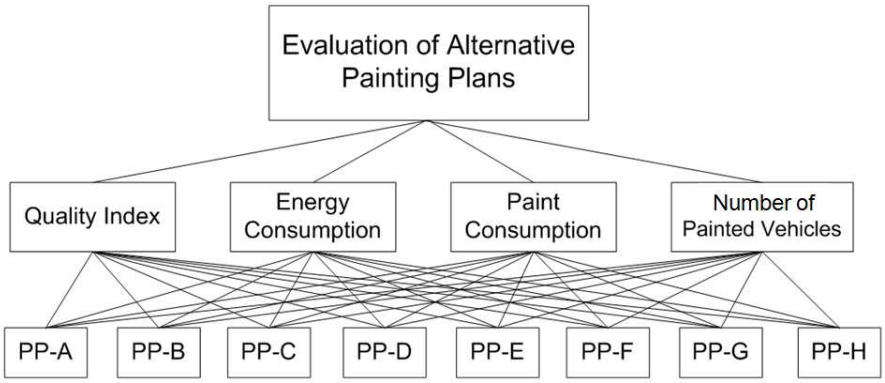


Figure 5.2: The tree hierarchy. The top level indicates the main goal of this MCDA process. The second level consists of the criteria that contribute to the overall goal. The third level comprises the options that will be evaluated in terms of the criteria of the second level.

are relevant in the sense that they indicate the degree of importance of each criterion to the overall goal. These weights are expressed in a ratio scale and the sum of the normalised weights equals to 1. Thus, each criterion can be interpreted according to their proportional importance.

The weights are assigned to the chosen criteria by asking the DM to compare the four criteria with each other in terms of their relative importance or contribution to the main goal of the decision problem, using the nine-point intensity scale proposed by Saaty (1977, 1986) (**Step 4**). Table 5.4 shows the pairwise comparison judgements provided by the DM, as well as the resulting criteria weights. Note that reciprocals are automatically assigned in each pairwise comparison.

Table 5.4: AHP pairwise comparison matrix for the chosen criteria and the corresponding criteria weights.

	QI	**EC**	**PC**	**NPV**	**Criteria Weights**
Quality Index (QI)	1	3	7	9	0.6055
Energy Consumption (EC)	$1/3$	1	2	7	0.2296
Paint Consumption (PC)	$1/7$	$1/2$	1	5	0.1255
Number of Painted Vehicles (NPV)	$1/9$	$1/7$	$1/5$	1	0.0394

Based on the AHP results, the Quality Index was deemed the most important criterion ($w_{QI} = 60.55\%$) for the evaluation of the painting plans, followed by Energy Consumption ($w_{EC} = 22.96\%$), and Paint Consumption ($w_{PC} = 12.55\%$). The least important criterion is the Number of Painted Vehicles, which was assigned a relative importance of merely 3.94%. This reveals the DM's preference for quality over quantity. The considerable weight assigned to the Quality Index criterion is consistent with the marketing idea that quality in automobile painting is one of the most important product features, being critical to product sales.

In order to ensure the consistency of the obtained criteria weights, we computed the *CR* for the pairwise comparison matrix presented in Table 5.4. Since $CR = 0.066 < 0.10$ was obtained, the DM has been coherent in his judgements and, thus, the criteria weights obtained can be used in the MCDA process.

5.4.2 PROMETHEE Computations

Additional information is required to proceed with the PROMETHEE method, namely the preference functions associated with each criterion and corresponding thresholds. This information was indirectly defined by the DM, based on his answers to a set of questions presented by the "Preference Function Assistant" of the Visual PROMETHEE software. These functions reflect the DM's perception of the criterion scale and are used to obtain the degrees of preference $P_j(a,b)$, $\forall a,b \in A$, where a and b are options pertaining to the finite set of options A, and j is the criterion index ($j = 1,...,n$). In face of quantitative criteria, the linear preference function as well as its special case (V-shape function) are usually the best choices. The selected preference functions and thresholds (absolute values) are provided in Table 5.5.

Table 5.5: Preference functions.

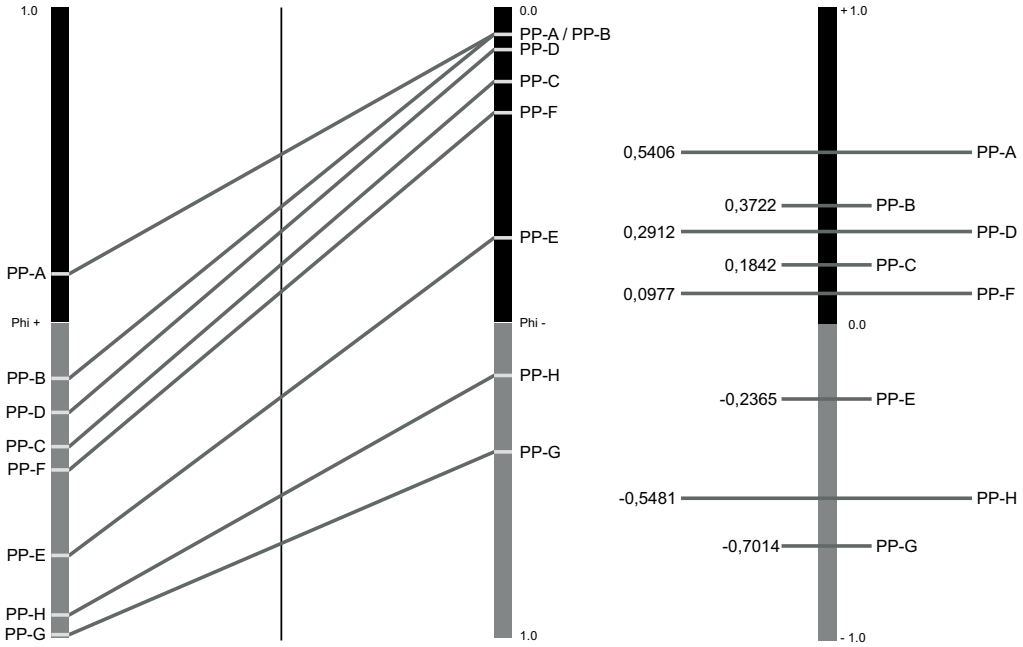
Criteria	Preference Function	Thresholds (absolute)	
		q	p
QI	V-Shape	-	0.95
EC	Linear	2	4
PC	V-Shape	-	0.31
NPV	Linear	2	4

The remaining steps are automatically performed by the Visual PROMETHEE software, without the need for human intervention (**Step 5**). After evaluating the options, the outranking flows (positive flow $\phi^+(\mathbf{a})$, negative flow $\phi^-(\mathbf{a})$, and net flow $\phi(\mathbf{a})$), displayed in Table 5.6, are outputted.

The PROMETHEE I ranks the options based on the comparison between the positive and negative flows (**Step 6**). The result is graphically presented on the left-hand side of Figure 5.3. In this figure, the left and right columns correspond to the $\phi^+(\mathbf{a})$ and $\phi^-(\mathbf{a})$ flows, respectively, and the middle column represents the balance between the positive and negative flows, thus corresponding to the net preference flow $\phi(\mathbf{a})$. This middle column is highlighted on the right-hand side of Figure 5.3 and represents the continuous scale of the net flows, from the highest (top) to the lowest (bottom) scores. The net flows are used to generate a complete ranking in the PROMETHEE II (**Step 7**). Options located closer to the top of these columns (left-hand side of Figure 5.3) have better performance. Options are deemed incomparable if they occupy different positions in the

Table 5.6: PROMETHEE outranking flows.

Options	ϕ^+	ϕ^-	ϕ
PP-A (Single + 1 Colour)	0.578	0.038	0.541
PP-B (Single + 2 Colours)	0.411	0.039	0.372
PP-C (Single + 3 Colours)	0.303	0.119	0.184
PP-D (Single + 4 Colours)	0.356	0.065	0.291
PP-E (Mixed + 1 Colour)	0.130	0.366	-0.237
PP-F (Mixed + 2 Colours)	0.266	0.169	0.098
PP-G (Mixed + 3 Colours)	0.006	0.707	-0.701
PP-H (Mixed + 4 Colours)	0.036	0.584	-0.548



(a) PROMETHEE I (b) PROMETHEE II

Figure 5.3: PROMETHEE I partial ranking (left-hand side) and PROMETHEE II complete ranking (right-hand side) for the paint shop problem under analysis.

ranking induced by each flow (*i.e.*, ($\phi^+(\mathbf{a})$ and $\phi^-(\mathbf{a})$). Since this is not the case, all options are comparable. It was also verified that there are no indifferent options since none of their flow scores ($\phi^+(\mathbf{a})$ and $\phi^-(\mathbf{a})$) overlap. As a consequence, both the PROMETHEE I and the PROMETHEE II return the same ranking, as can be ascertained from the analysis of both sides of Figure 5.3.

The analysis of both rankings for the paint shop problem indicates that the PP-A is the most preferred option. In other words, the painting plan with the highest relative merit is the one involving the painting of single cabins with only one colour. This result is expected since PP-A achieves the best value in the two most important criteria, namely QI and EC. From the business point of view, this result means that the PP-A is the most satisfactory compromise solution since it corresponds to the painting plan that contributes the most to the paint shop optimisation. The other options are ranked in the following order: PP-B, PP-D, PP-C, PP-F, PP-E, PP-H, and PP-G (please see Table 5.7). It is also clear from Figure 5.3 that PP-B, PP-D, PP-C, and PP-F are close to each other in the upper-middle of the PROMETHEE II ranking, with low differences in their $\phi(\mathbf{a})$ scores. Painting plan PP-G, which involves the painting of mixed cabins with three distinct colours, is the least preferred option, appearing at the bottom of the ranking.

Table 5.7: Final ranking yielded by the PROMETHEE I and PROMETHEE II methods. In our application, the rankings obtained for each method coincide.

Painting Plans	Description
PP-A	Single Cabin + 1 Colour
PP-B	Single Cabin + 2 Colours
PP-D	Single Cabin + 4 Colours
PP-C	Single Cabin + 3 Colours
PP-F	Mixed Cabin + 2 Colours
PP-E	Mixed Cabin + 1 Colour
PP-H	Mixed Cabin + 4 Colours
PP-G	Mixed Cabin + 3 Colours

In order to visually understand the obtained results, the multicriteria problem has been represented in the GAIA plane (**Step 8**), see Figure 5.4. Here, options are represented by squared points and criteria are represented by vectors. Criteria expressing similar preferences are oriented in the same direction, whereas conflicting criteria point into different directions. The length of the criterion vector is an indicator of its discriminative power. Hence, longer vectors are associated with criteria that are more effective differentiating options. The quality level of the GAIA projection is given by the Δ-parameter. The GAIA plane of Figure 5.4 has a Δ-parameter of 89.8%, which means that only 10.2% of the total information gets lost by the projection of a four-dimensional criteria

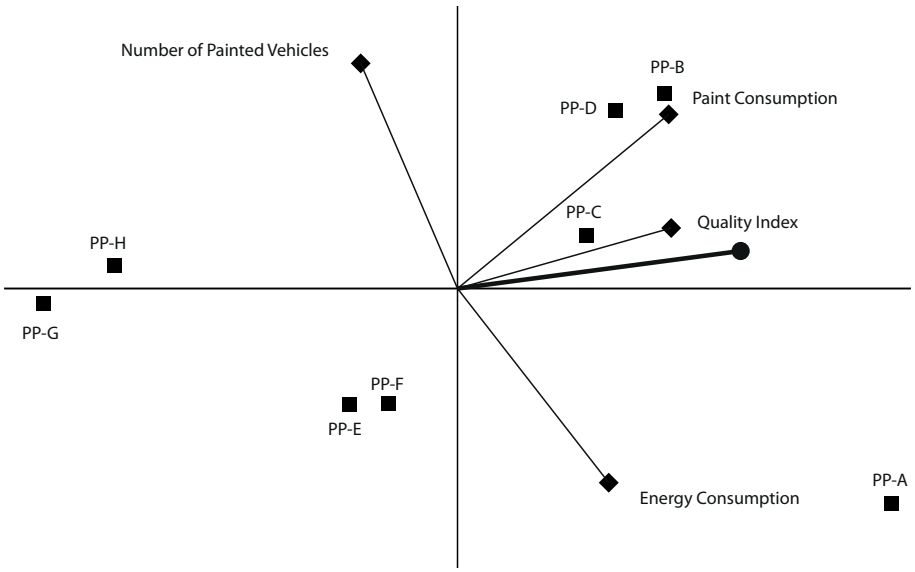


Figure 5.4: GAIA plane. The Δ-parameter for this visualisation is 89.8%, which means that a satisfying amount of the total information was preserved in the two-dimensional projection.

space onto a two-dimensional space yielded by a Principal Component Analysis. The analysis of this figure shows that the differentiating power of the EC, PC, and NPV criteria is identical, as indicated by the similar length of the corresponding vectors. The criterion QI has a slightly lower ability to discriminate among options. Regarding QI, it can also be observed that this criterion expresses a preference similar to PC, since they have closer orientations. In this study, this means that we can find options having good scores in both QI and PC (*e.g.*, PP-A, PP-B, PP-D). The GAIA plane also reveals the strong conflicting nature of two criteria: EC and NPV. This result makes sense in the context of our case study, since increased production of vehicles typically implies higher levels of energy consumption. The GAIA plane also provides information concerning the performance of the painting plans with regard to the different criteria. For instance, option PP-A is especially good on the Energy Consumption criterion, whereas PP-B is the best option in terms of Paint Consumption. Groups of solutions with similar profiles can also be identified in the plane, since they tend to cluster together (*e.g.*, options PP-G and PP-H have a similar profile). Information regarding criteria weights is also given in Figure 5.4 through vector pi (also called decision axis). Vector pi is depicted as a circle-ended thick line and can be interpreted as a weighted average of the criteria axes. It represents the direction of the compromise, when taking into account the DM's perceived relative importance of criteria (*i.e.*, the criteria weights). The amplitude obtained by projecting the options onto the decision axis reveals which are the most promising options. More specifically, options located in the rightmost side of the decision axis score higher on the most important criteria and, thus, represent a better compromise solution, whereas options located closer to the origin of the decision axis are less promising. Using this reasoning, Figure 5.4 reveals that the

best compromise solution is the PP-A, followed by PP-B (this is consistent with the PROMETHEE I and PROMETHEE II rankings, as expected).

Table 5.8: Weight stability intervals, for the AHP-PROMETHEE (left-hand side) and for the PROMETHEE (right-hand side), when taking into account the entire set of options (full ranking). The largest interval amplitude for each criterion, when comparing both methods, is underlined.

AHP-PROMETHEE					**PROMETHEE**			
Criteria	**Weights**	**Stability Interval**			**Weights**	**Stability Interval**		
	(AHP)	Min	Max	Amplitude	**(Swing-Weight)**	Min	Max	Amplitude
QI	0.6055	0.355	0.821	<u>0.466</u>	0.5880	0.351	0.805	0.454
EC	0.2296	0.033	0.411	<u>0.378</u>	0.2350	0.064	0.410	0.346
PC	0.1255	0.043	0.539	<u>0.496</u>	0.1180	0.041	0.533	0.492
NPV	0.0394	0	0.17	0.17	0.0590	0	0.176	<u>0.176</u>

Since some steps of the MCDA process can be permeated by subjectivity and uncertainty, a sensitivity analysis was conducted in an attempt to understand the magnitude of the impact of different criteria weighing schemes on the PROMETHEE II complete ranking. This was performed by identifying stability intervals for criteria weights (**Step 9**). These intervals provide the range in which the criteria weights can vary without changing the complete ranking, thus allowing to infer on the robustness of the method. The higher the amplitude of the interval is, the more robust the method is to changes in criteria weights. The amplitude of these intervals also makes it easy to identify the most sensitive criterion, which are the ones associated with smaller amplitudes. Table 5.8 provides the stability intervals for the criteria, when taking into account the entire set of options (full ranking), for two different scenarios: the application of PROMETHEE using the criteria weights yielded by the AHP (*i.e.*, the AHP-PROMETHEE), and the application of PROMETHEE using the criteria weights obtained by the swing-weight procedure (Von Winterfeldt and Edwards, 1986) (*i.e.*, the PROMETHEE). Details for the second scenario will be given further along in the chapter. From the analysis of the stability intervals, we conclude that the NPV is the less robust criterion since its corresponding stability interval has the lowest range. Thus, a significant change in the NPV weight may modify the position of the options in the complete ranking. The remaining criteria (EC, QI, and PC) are more robust since their weights can vary over a very larger range. The comparison of the amplitudes of the AHP-PROMETHEE stability intervals with the ones obtained for the PROMETHEE (Table 5.8) shows that the former is more robust since its interval amplitudes are larger in three, out of the four criteria considered. A similar analysis was performed only for the first-ranked option (PP-A) aiming to assess the robustness of its position on the complete ranking. The results are reported in Table 5.9. As can be ascertained from this table, the position of PP-A at the top of both the AHP-PROMETHEE and the PROMETHEE ranking is robust for three criteria (QI, EC, and PC), but sensitive to changes in the weight of the NPV criterion. For instance, if the weight assigned to NPV were 20%, PP-B would replace PP-A in the ranking. In turn, QI is a very

robust criterion regarding the ranking position of PP-A, as indicated by the maximum amplitude of its weight stability interval, for both methods. Again, the results show the AHP-PROMETHEE to be more robust than the PROMETHEE, although not by far.

Table 5.9: Weight stability intervals for the first-ranked option, when using two different methods: the AHP-PROMETHEE (left-hand side) and the PROMETHEE (right-hand side). The largest interval amplitude for each criterion, when comparing both methods, is underlined.

AHP-PROMETHEE					**PROMETHEE**			
Criteria	**Weights**	**Stability Interval**			**Weights**	**Stability Interval**		
	(AHP)	Min	Max	Amplitude	**(Swing-Weight)**	Min	Max	Amplitude
QI	0.6055	0	1	1	0.5880	0	1	1
EC	0.2296	0.033	1	<u>0.967</u>	0.2350	0.064	1	0.936
PC	0.1255	0	0.784	<u>0.784</u>	0.1180	0	0.764	0.764
NPV	0.0394	0	0.17	0.17	0.0590	0	0.176	<u>0.176</u>

In the scope of the application, and for comparison purposes, the same problem has also been analysed with the stand-alone versions of the AHP and the PROMETHEE methods. The results of the full AHP are obtained by applying the complete procedure described in Section 5.3.1. The full application involved an intermediate step, namely, the elicitation of preferences for the different painting plans according to their contribution to each criterion, which required additional meetings with the DM. Having the information regarding both the criteria weights, derived in Step 4 of the proposed methodology, and the priorities for the options, we carried out a weighing and summing step that yielded the global priorities of the options. Regarding the PROMETHEE, since the proposed methodology heavily relies on this method, the only difference between them is the adopted procedure to elicit the criteria priorities. Instead of using the AHP method for this step, we followed the swing-weight procedure proposed by Von Winterfeldt and Edwards (1986). According to this procedure, the DM should first identify the most important criterion, to which a score of 100 is assigned, and then successively allot relative scores (lower than 100) to the second, third, and fourth most important criteria. The given scores should reflect the DM's priorities and perceptions, as well as the magnitude of preference. These numbers are further divided by their sum so as to obtain the normalised criteria weights. The obtained criteria weights were: $w_{QI} = 58.8\%$, $w_{EC} = 23.5\%$, $w_{PC} = 11.8\%$, and $w_{NPV} = 5.9\%$. The order of criteria importance returned by the swing-weight procedure is consistent with the one derived using AHP, QI being the criterion with highest relative importance and the NPV the criterion deemed less important for the evaluation of painting plans. One caveat regarding these results: although these tasks have been performed at distinct moments in time, the criteria weights obtained by the swing-weight procedure are potentially biased by the previous application of the AHP to elicit criteria weights.

The results reported in Table 5.10 show that the three MCDA methods produce the same ranking. This was expected for both versions of the PROMETHEE, despite the distinct approaches

Table 5.10: Rankings returned by three MCDA methods: the AHP, PROMETHEE, and AHP-PROMETHEE methods. The first row corresponds to the most preferred option and the last row to the least preferred option.

AHP		**PROMETHEE**		**AHP-PROMETHEE**	
Ranking	Global Priorities	Ranking	Net Flows	Ranking	Net Flows
PP-A	36.54	PP-A	0.51	PP-A	0.54
PP-B	15.88	PP-B	0.36	PP-B	0.37
PP-D	14.03	PP-D	0.28	PP-D	0.29
PP-C	11.83	PP-C	0.18	PP-C	0.18
PP-F	10.88	PP-F	0.10	PP-F	0.10
PP-E	6.61	PP-E	-0.23	PP-E	-0.24
PP-H	2.46	PP-H	-0.53	PP-H	-0.55
PP-G	1.77	PP-G	-0.68	PP-G	-0.7

to elicit criteria weights, because the weights generated by the swing-weight procedure are within the stability intervals of the AHP-PROMETHEE (*cf.* left hand-side of Table 5.8). The agreement between the three MCDA methods shows the reliability of the methodological approach used in this research (the AHP-PROMETHEE) and strengthens the DM's confidence in the obtained ranking. Even though these methods lead to the same results, they are not replaceable. We consider that the combination of the modelling and problem formulation and structuring of the AHP with the user friendliness, simplicity of the model strategy, and implementation of the PROMETHEE, makes the hybrid AHP-PROMETHEE approach more appealing than the stand-alone methods (the AHP and the PROMETHEE), from the application point of view.

5.5 Conclusions and Future Work

This chapter proposes the use of a hybrid approach that combines the strengths of the well known AHP and PROMETHEE methods to support the decision analysis process in a paint shop of an automobile assembly plant. The goal was to assist the management in the process of evaluating the relative merits of the most frequently used painting plans, in order to enhance the paint shop planning and the optimisation of the painting activities. This problem is of great relevance for the company, since the paint shop accounts for the highest costs of the plant and is a bottleneck of the assembly line. The application of this integrated approach to the paint shop decision problem was presented in detail by traversing each of the steps of the AHP-PROMETHEE methodology.

The AHP-PROMETHEE relies on the idea that there are a number of distinctive characteristics of the AHP that could be used to enhance the PROMETHEE, namely at the level of structuring the decision problem and at the stage of determining criteria weights. The AHP hierarchy tree can improve the PROMETHEE since in its original formulation the latter did not consider the possibility of structuring the decision problem. The lack of a problem structuring phase can be seen

as a disadvantage, especially when there are many levels of criteria, since it might preclude the decision maker from obtaining a clear view of the decision problem at hand. Another advantage of the AHP is the use of a formal and systematic procedure for eliciting criteria weights, which relies on pairwise comparisons and on a semantic scale to represent trade-offs between criteria. The incorporation of the AHP weights, yielded by this procedure, into the PROMETHEE, lessens the ambiguity associated with the process of weighing criteria and increases the rigour and transparency of the MCDA process. In turn, the PROMETHEE enriches the AHP by associating a preference function with each criterion and by offering a set of tools (*e.g.*, the GAIA plane) to thoroughly analyse the decision problem.

The case study in the paint shop of an automobile assembly plant showed that the AHP-PROMETHEE is operationally useful, being easy to apply and easy to understand by non-experts (*i.e.*, the DM). By comparing the results of the proposed method with the ones obtained using the stand-alone versions of the AHP and PROMETHEE, it was also empirically demonstrated the added reliability and robustness of the AHP-PROMETHEE approach. These results validate the hybrid methodological approach and, therefore, the obtained ranking. Regarding the results of the AHP-PROMETHEE on the evaluation of the different painting plans, based on the used data sample it was found that the most preferred compromise solution corresponds to the daily painting plan PP-A (painting of single cabins with a single colour). Nevertheless, due to the type of sampling used to collect the data, the statistical characteristics of the data, and the uniqueness of each MCDA process, the empirical results of this study cannot be generalised to other assembly plants. In what regards the decision support, the management found the logic behind the proposed approach easy to understand, the interpretation of the required inputs objective, and the results of the evaluation interesting and of value. On a higher level, the application of MCDA was advantageous to the DM since it stirred reflection on the decision problem, allowing a more structured analysis of its characteristics and consequences. As a result, the plant is already making use of the final ranking in the weekly scheduling of the paint shop. The release of production capacity made possible by the increased operational efficiency of the painting activities, achieved with the results of this research, allowed this plant to won the bid to start assembling a new automobile model.

We acknowledge that this study is limited in ways that suggest opportunities for future research. A possible direction would be to incorporate in the model the vagueness and uncertainty associated with the DM's judgements by using the fuzzy versions of the AHP and the PROMETHEE.

Part IV

Conclusions

Chapter 6

Concluding Remarks

6.1 Main Conclusions

Contemporary organisations produce and have access to sheer amounts of data, facing the so-called data deluge problem. As data become huge, more current and fine-grained, decision makers find it hard to manually look for and identify relevant pieces of information, which are needed to inform and support their decisions. The ubiquitous availability of data and decreasing costs in storage and computational power, coupled with increasing market pressures created favourable conditions for the emergence of an analytics-driven management in modern organisations. This type of management is characterised by a strong emphasis in data-driven decision making, since it values decisions that can be backed up with data and facts, and is fuelled by the widespread presence and use of Business Analytics solutions across several organisational levels. Business Analytics methods and tools can handle large volumes of different types of data and are able to distill what is valuable in these data, with the ultimate purpose of delivering relevant insights to managers, in a timely manner. Companies that embrace Business Analytics are likely to achieve benefits ranging from increased revenue to lowered costs, reduced risk, enhanced processes, better products and services, and improved customer loyalty and satisfaction. Thus, the adoption and effective use of Business Analytics in organisations can act as a competitive differentiator in the market.

Motivated by this context, the research presented in this dissertation aimed to develop and apply new analytical approaches that help organisations exploit the historical data they have been accumulating over time. In particular, our goals were to (i) develop new data-driven methodologies for advanced descriptive analytics that are able to extract and visualise temporal and structural patterns from time-evolving network data, and to (ii) investigate the potential of using hybrid multicriteria decision analysis methods to address decision problems in complex industrial systems. These

research goals were further decomposed into three research questions. Based on the empirical investigation described in the preceding chapters, it is now possible to answer them:

Descriptive Analytics:

RQ1: *How can we model and detect changes in historical data by harnessing both the relational and the temporal nature of data?* We model time-stamped relational data as an evolving network, *i.e.*, as a temporally ordered sequence of mathematical graphs, each depicting the state of the network at a given time step. Nodes in this network represent entities, such as customers or products. These nodes are connected to each other by links that represent the relationships, or interactions, established between them (co-purchasing, similarity, formal or informal communication, flow of information). Information regarding the direction, intensity or frequency of these relationships, is introduced in the network model by assigning a direction and weights to the links. The use of a network model to represent historical relational data is advantageous since it has a solid mathematical foundation and, at the same time, enables the exploration and application of a wide diversity of methods and techniques from social network analysis and network science. This network model was explored in Chapter 3. We also model evolving networks as three-order tensors, where rows represent the nodes, columns represent node centrality measures and fibers represent time. This modelling strategy was explored in Chapter 4. Regarding the analysis of changes in evolving networks, this can be performed at three levels of analysis: the network-level, the community-level and the node-level. At the network-level, changes are detected through the analysis of time series of several characteristics of the network (degree distribution, density, effective diameter, average path length, size of the connected components, among others) using SNA measures and techniques. A similar procedure can be used to identify changes in the evolution of nodes, by studying the time series of their centrality measures. At the community-level, changes are detected using methods for dynamic community evolution analysis. In Chapter 3 we propose the use of an event-based framework, dubbed MECnet, to perform this task. MECnet comprises two independent stages: community discovery and community monitoring. In the first stage, densely connected groups of nodes (the communities) are independently discovered at each snapshot of an evolving network, so as to uncover the community structure of the network at consecutive time steps. The second stage encompasses a sequence of steps that aim at (i) finding dynamic communities through community matching, (ii) detecting critical evolutionary events in the evolution of these dynamic communities (community births, survivals, splits, merges and deaths) and, based on this, (iii) characterise their life-cycle. The outputs of MECnet are easy to grasp and understand by non-experts and provide a high-level summary of the temporal changes taking place in the evolving network.

RQ2: *How can we visualise and understand the relational meaning of the detected temporal changes?* Understanding possible reasons behind temporal changes in evolving networks is a key step towards insight. Besides detecting when a given type of change occurred, it is important to express the detected change in a clear, comprehensible way and advance possible reasons for why it happened. Moving the focus from the "What?" and "When?" to the "Why?" is essential to obtain a more complete view of historical data. Since graphic visualisation is often more effective at conveying complex patterns than text or numbers, in Chapter 4 we introduce a novel visualisation approach that provides a spatial context to the temporal evolution of dynamic networks, by shedding light on the network's structural changes. The proposed visualisation relies on displaying spatio-temporal trajectories of nodes, or communities, in a compact and interpretable 2D space, whose axes represent latent relational properties of the network derived from network-level and node-level centrality measures. This space is generated based on the output of a Tucker3 model with orthogonality constraints, which takes as input a three-order tensor embedded with structural information derived from the network. The spatio-temporal trajectories are projected in the 2D space and map the temporal evolution of a node or community. The spatial movement of a given trajectory in this space reveals not only temporal changes but also associates them with changes in the network structure. These structural changes are interpreted according to the latent properties captured in the axes of the 2D space and provide a relational meaning to the temporal changes.

Prescriptive Analytics:

RQ3: *What is the potential of using hybrid approaches in multicriteria decision analysis to aid decision makers address real-life complex decision problems?* The integration of different multicriteria decision analysis methods to address decision problems in industrial complex systems seems to be a promising approach in the context of prescriptive analytics, as suggested by the empirical results of our case study in one of Toyota's assembly plants (Chapter 5). In this case study, we applied the AHP-PROMETHEE hybrid approach, aided by GAIA, to the novel problem of evaluating the relative merits of the most frequently used daily painting plans, with the goal of helping the decision maker enhance the paint shop planning and optimise the painting process. Based on this applied research, we concluded that: (i) the AHP-PROMETHEE hybrid approach is operationally useful, being easy to apply and easy to understand by non-experts; (ii) the AHP-PROMETHEE approach is slightly more reliable and robust than the corresponding stand-alone versions, and (iii) the GAIA plane is a powerful visual tool that plays an important role in the dissemination of the MCDA

results to the decision maker, since it allows for a comprehensive analysis of the multicriteria problem, revealing not only the recommended options but also why they are recommended. Although the final decision is up to the decision maker, our results suggest that the use of hybrid approaches for addressing multicriteria decision problems can provide a reliable path to the resolution of operational problems. Besides, the flexibility of the MCDA process is increased with the integration of two or more MCDA methods, since it allows the analysts to explore and apply the combination of methods that best suit the decision problem at hand.

6.2 Main Contributions

The empirical research conducted in this doctoral thesis produced a series of original contributions to the fields of business analytics, social network analysis, dynamic network analysis, information visualisation and multicriteria decision analysis. In Part II of the thesis, our scientific contributions were mainly methodological, as we devised novel methodologies for performing descriptive analytics over dynamic network data. In Part III of the thesis, we contribute to the field of business analytics and multicriteria decision analysis by applying an existing method (the AHP-PROMETHEE) to a new decision problem and showing the suitability of this hybrid approach in a new context.

The methodologies proposed within the scope of descriptive analytics contribute to the Business Analytics field as they explore advanced methods and techniques for handling, analysing and visualising network data (*i.e.*, data that describes the relationships among entities) from a temporal perspective. Our focus on the dynamic and relational nature of data allows managers to detect patterns and events which would go unnoticed if the analysis was performed on data aggregated over the entities' attributes, which is the traditional approach. We also contribute to the prescriptive analytics perspective of Business Analytics by applying a hybrid approach that integrates the strengths of the well known AHP and PROMETHEE methods to a novel multicriteria decision problem in the paint shop of an automobile assembly plant. Our case study in one of Toyota's assembly plants provides a detailed description on how multicriteria decision analysis can be used to tackle the new problem of evaluating vehicle painting plans, with the goal of enhancing the paint shop planning and the optimisation of the painting activities. This case study represents an example of the application of prescriptive analytics in a complex industrial system, in which a ranking of several decision options is provided to the decision maker and their likely outcomes are made explicit.

6.3 Directions for Future Research

The new methodologies described in this dissertation can be improved and extended in a number of ways, opening several opportunities for future research. Some ideas are outlined below.

- Extension of our methodology for monitoring the evolution of customer profiles by incorporating a prediction step. An idea would be to predict the most likely event the dynamic community will experience in the next time step. This would imply not only the analysis of its past sequence of critical events but also the analysis of its structural proximity to other dynamic communities in the network. This prediction task could eventually support the development of a recommendation system that suggests the most likely product a customer will buy based on his/her community membership;

- Application of the previous methodology to large-scale networks collected in different types of industries and markets (*e.g.*, banking, telecommunications, retail, web commerce), in order to evaluate the feasibility of the methodology in new contexts and assess the relevance of the extracted patterns in providing insights into customer behaviour;

- Application of our visualisation methodology in real-world business problems, such as churn prediction and viral marketing.

- Creation of a tool that integrates both methodologies, so as to facilitate and promote their use in organisations and in science.

Appendices

Appendix A

Tucker3 Model Outputs

This Appendix presents the rotated component matrices **A**, **B** and **C** yielded by the application of an orthogonality-constrained Tucker3 model $(4 \times 3 \times 4)$ to a three-order tensor embedded with structural information from an evolving friendship network described in Chapter 4.

Table A.1: Rotated component matrix **B** resulting from the application of a Tucker3 model of order $(4 \times 3 \times 4)$ to the undirected binary version of Van de Bunt's temporal friendship network. B1, B2 and B3 refer to the first, second and third components of matrix **B**. The largest absolute coefficient for each node-level measure is highlighted in bold. The last column of the matrix gives the partitioned fit proportion for each one of the node-level measures. The fit proportion is obtained by dividing the entity's fit by the total sum of squares of the data entries associated with the corresponding entity. Large fit values indicate that the entity is well represented in the model, whereas poor fit values suggest the presence of noise in the associated data elements or an incomplete modelling of these data.

	B1	**B2**	**B3**	**Fit**
Degree	−0.11	**0.53**	0.31	86.51%
Eigenvector Centrality	**0.86**	−0.18	0.33	75.01%
Closeness Centrality	0.07	0.02	**−0.78**	94.38%
Betweenness Centrality	−0.12	**0.58**	0.22	87.39%
Clustering Coefficient	−0.48	**−0.59**	0.36	69.94%

Table A.2: Rotated component matrix **C** resulting from the application of a Tucker3 model of order $(4 \times 3 \times 4)$ to the undirected binary version of Van de Bunt's temporal friendship network. C1, C2, C3 and C4 refer to the first, second, third and fourth components of matrix **C**. The largest absolute coefficient for each timestep/occasion is highlighted in bold. The last column of the matrix gives the partitioned fit proportion for each one of timesteps.

	C1	**C2**	**C3**	**C4**	**Fit**
t_0	0.05	0.03	0.02	**1**	99.80%
t_1	0.28	0.43	**−0.49**	0.04	81.89%
t_2	−0.1	0.47	**−0.49**	0	86.39%
t_3	**0.69**	0.3	0.32	−0.05	55.70%
t_4	0.2	**0.39**	0.34	0.01	69.83%
t_5	**−0.58**	0.5	0.11	0.02	88.45%
t_6	−0.25	0.31	**0.53**	−0.04	57.14%

Table A.3: Rotated component matrix **A** resulting from the application of a Tucker3 model of order $(4 \times 3 \times 4)$ to the undirected binary version of Van de Bunt's temporal friendship network. A1, A2, A3 and A4 refer to the first, second, third and fourth components of matrix **A**. Students are referred to as S1, S2, S3 through to S32. The largest absolute coefficient for each student is highlighted in bold (except in those cases where the coefficient is very low or very similar acrosss components). The last column of the matrix gives the partitioned fit proportion for each one of the entities of mode *A* (the students).

	A1	**A2**	**A3**	**A4**	**Fit**
S1	−0.01	−0.01	−0.03	0.04	6.51%
S2	0.02	**0.48**	0.02	0.15	92.04%
S3	0.3	−0.08	0.12	**0.18**	75.6%
S4	−0.15	−0.09	0.09	**0.24**	71.11%
S5	0	0	0.01	0	0.99%
S6	−0.01	−0.01	−0.03	0.04	7.87%
S7	0.27	−0.02	**0.29**	0.07	75.74%
S8	0.06	−0.11	−0.3	**0.32**	89.46%
S9	0.02	−0.01	0	0.02	8.81%
S10	0.2	−0.03	**0.26**	0.13	40.50%
S11	−0.02	−0.01	−0.04	0.04	8.30%
S12	0.17	−0.1	**−0.27**	0.23	71.21%
S13	0.03	−0.06	0.12	**0.15**	55.78%
S14	0.09	−0.03	−0.04	**0.11**	30.63%
S15	0.15	−0.03	**0.26**	0.11	64.49%
S16	0.16	**0.49**	0.23	0.09	94.73%
S17	−0.08	**0.45**	−0.15	0.24	98.82%
S18	0	0	0	0	0%
S19	**−0.2**	−0.07	0.04	0.18	81.67%
S20	−0.29	−0.06	**0.32**	0.17	89.04%
S21	0.07	−0.03	−0.06	**0.11**	35.88%
S22	0.1	−0.1	−0.27	**0.29**	84.28%
S23	**0.27**	−0.08	0.12	0.19	82.37%
S24	**0.31**	−0.1	−0.05	0.25	85.81%
S25	0.13	−0.05	0.1	**0.15**	47.53%
S26	**0.2**	−0.08	**−0.2**	0.17	71.64%
S27	**−0.25**	−0.08	0.19	0.22	85.66%
S28	**−0.36**	−0.11	−0.07	0.31	91.41%
S29	**−0.32**	−0.1	0	0.28	95.63%
S30	−0.13	−0.04	**0.37**	0.14	84.89%
S31	−0.08	**0.46**	−0.11	0.2	95.64%
S32	0.08	−0.03	**0.25**	0.09	73.2%

Appendix B

Statistical Analysis of the Paint Shop Data

This Appendix reports some descriptive statistics of the paint shop data sample made available by Toyota. This data sample was used for conducting the study reported in Chapter 5 (Tables B.1 and B.2). The following statistics were computed and analysed in order to appraise the suitability of this data sample to our application:

- *Arithmetic mean* μ and *median* $\widetilde{x}$
- *Variance* σ^2 and *standard deviation* σ
- *Coefficient of variation CV* (Equation B.1): we compute this measure to assess the validity of using the available data to address Toyota's multicriteria problem. The coefficient of variation is a standardised measure of dispersion, which indicates the degree of data variability with respect to the mean.
- *Pearson's second skewness coefficient* Sk_2 (Equation B.2): we compute this coefficient to ascertain the validity of using the mean, instead of the median, as a measure of central tendency of the data sample. The results were analysed as follows: if the distributions are very asymmetric (large positive or large negative values of Sk_2), the median is more appropriate than the mean to represent the data, due to its robustness against outliers and against extremely high or low data values. If the distribution is symmetric (*i.e.*, $Sk_2 = 0$), the mean is equal to the median, thus it is indifferent to use the mean or the median to represent the paint shop data sample.
- *Pearson's correlation coefficient* $\rho_{X,Y}$ (Equation B.3): since the criteria is quantitative, we compute the Pearson's correlation coefficient for every pairwise combination of criteria in order to appraise the validity of the assumption of the preferential independence of pairs of

criteria. To ensure that this assumption is not violated, the correlation values should be close to zero.

$$\text{Coefficient of Variation: } CV = \frac{\sigma}{\mu} \tag{B.1}$$

$$\text{Pearson's second skewness coefficient: } Sk_2 = \frac{\mu - \widetilde{x}}{\sigma} \tag{B.2}$$

$$\text{Pearson's correlation coefficient: } \rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{B.3}$$

Table B.1: Criteria correlation matrix for the paint shop data sample.

	QI	**EC**	**PC**	**NPV**
QI	1	0.02	0.3	0.07
EC	0.02	1	0.07	0.36
PC	0.3	0.07	1	-0.14
NPV	0.07	0.36	-0.14	1

Table B.2: Descriptive statistics of the paint shop data sample made available by Toyota, for each painting plan.

Alternatives/Criteria		QI	EC	PC	NPV
Unit		# Defects	kWh	Ink liters	# Painted vehicles
Max/Min		Min	Min	Min	Max
PP-A	Mean	2.14	18	1.59	10
	Median	2	19	1.61	9
	Std. Deviation	0.77	3.27	0.18	4.79
	Variance	0.59	10.68	0.03	22.98
	Pearson's Skewness Coefficient	0.54	-0.84	-0.28	0.9
	Coefficient of Variation	0.36	0.18	0.11	0.46
PP-B	Mean	2.19	23	1.57	14
	Median	2.15	22	1.43	14
	Std. Deviation	0.65	4.27	0.15	5.99
	Variance	0.42	18.22	0.02	35.85
	Pearson's Skewness Coefficient	0.17	0.53	2.89	-0.02
	Coefficient of Variation	0.3	0.19	0.09	0.43
PP-C	Mean	2.58	21	1.64	14
	Median	2.4	21	1.44	14
	Std. Deviation	0.91	4.04	0.2	2.43
	Variance	0.82	16.32	0.04	5.92
	Pearson's Skewness Coefficient	0.61	-0.41	3.09	-0.47
	Coefficient of Variation	0.35	0.19	0.12	0.18
PP-D	Mean	2.33	23	1.6	16
	Median	2.1	22	1.54	15
	Std. Deviation	0.68	1.35	0.27	2.31
	Variance	0.46	1.82	0.07	5.33
	Pearson's Skewness Coefficient	1.03	0.35	0.62	1.73
	Coefficient of Variation	0.29	0.06	0.17	0.14
PP-E	Mean	3	21	1.81	13
	Median	2.6	22	1.79	14
	Std. Deviation	1.56	3.99	0.3	2.34
	Variance	2.43	15.89	0.09	5.47
	Pearson's Skewness Coefficient	0.77	-0.71	0.17	-0.81
	Coefficient of Variation	0.52	0.19	0.16	0.17
PP-F	Mean	2.52	21	1.92	14
	Median	2.3	22	1.73	14
	Std. Deviation	0.74	4.1	0.15	3.14
	Variance	0.55	16.79	0.02	9.85
	Pearson's Skewness Coefficient	0.88	-0.98	3.78	-0.12
	Coefficient of Variation	0.3	0.2	0.08	0.23
PP-G	Mean	3.45	25	1.92	15
	Median	2.85	25	1.83	15
	Std. Deviation	1.55	2.05	0.24	3.06
	Variance	2.41	4.18	0.06	9.37
	Pearson's Skewness Coefficient	1.16	0.27	1.12	0.65
	Coefficient of Variation	0.45	0.08	0.13	0.2
PP-H	Mean	2.95	25	1.74	17
	Median	2.85	25	1.83	15
	Std. Deviation	0.99	3.68	0.35	2.06
	Variance	0.98	13.58	0.12	4.25
	Pearson's Skewness Coefficient	0.76	-0.08	1.2	-0.36
	Coefficient of Variation	0.31	0.15	0.18	0.13

Bibliography

Acar, E. and Yener, B. (2009). Unsupervised multiway data analysis: a literature survey. *IEEE Transactions On Knowledge and Data Engineering*, 21(1):6–20.

Acito, F. and Khatri, V. (2014). Business analytics: Why now and what next? *Business Horizons*, 57(5):565–570.

Aggarwal, C. and Subbian, K. (2014). Evolutionary network analysis: A survey. *ACM Computing Surveys (CSUR)*, 47(1):10.

Agha, S. R., Nofal, L. G., and Nassar, H. A. (2012). Multi-criteria governmental crop planning problem based on an integrated AHP-PROMETHEE approach. *International Journal of Applied Management Science*, 4(4):385–406.

Ahn, J.-w., Plaisant, C., and Shneiderman, B. (2011). A task taxonomy for network evolution analysis. Human-Computer Interaction Lab Tech Report HCIL-2011-09, University of Maryland.

Aigner, W., Miksch, S., Müller, W., Schumann, H., and Tominski, C. (2007). Visualizing time-oriented data - a systematic view. *Computers & Graphics*, 31(3):401–409.

Albert, R., Jeong, H., and Barabási, A.-L. (1999). Internet: Diameter of the World-Wide Web. *Nature*, 401(6749):130–131.

Allison, S. T., Jordan, A. M. R., and Yeatts, C. E. (1992). A cluster-analytic approach toward identifying the structure and content of human decision making. *Human Relations*, 45(1):49–72.

Alon, U. (2003). Biological networks: the tinkerer as an engineer. *Science*, 301(5641):1866–1867.

Andrade Jr, J., Bezerra, D., Ribeiro Filho, J., and Moreira, A. (2006). The complex topology of chemical plants. *Physica A: Statistical Mechanics and its Applications*, 360(2):637–643.

Archambault, D., Abello, J., Kennedy, J., Kobourov, S., Ma, K.-L., Miksch, S., Muelder, C., and Telea, A. C. (2014). Temporal multivariate networks. In Kerren, A., Purchase, H. C., and Ward, M. O., editors, *Multivariate Network Visualization*, volume 8380 of *Lecture Notes in Computer Science*, pages 151–174. Springer.

Arrow, K. J. (1951). *Social choice and individual values*. Wiley New York.

Asur, S., Parthasarathy, S., and Ucar, D. (2009). An event-based framework for characterizing the evolutionary behavior of interaction graphs. *ACM Transactions on Knowledge Discovery from Data*, 3(4):1–36.

Avikal, S., Mishra, P., and Jain, R. (2014). A fuzzy AHP and PROMETHEE method-based heuristic for disassembly line balancing problems. *International Journal of Production Research*, 52(5):1306–1317.

Avikal, S., Mishra, P. K., and Jain, R. (2013). An AHP and PROMETHEE methods-based environment friendly heuristic for disassembly line balancing problems. *Interdisciplinary Environmental Review*, 14(1):69–85.

Aynaud, T. and Guillaume, J.-L. (2010). Static community detection algorithms for evolving networks. In *Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, pages 513–519. IEEE.

Bach, B., Pietriga, E., and Fekete, J.-D. (2014). Visualizing dynamic networks with matrix cubes. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pages 877–886. ACM.

Bader, B. W., Harshman, R. A., and Kolda, T. G. (2006). Temporal analysis of social networks using three-way DEDICOM. Technical Report TR SAND2006-2161, Sandia National Laboratories.

Badrí, M. A. (2001). A combined AHP-GP model for quality control systems. *International Journal of Production Economics*, 72(1):27–40.

Bana e Costa, C., Ensslin, L., Cornêa, É. C., and Vansnick, J.-C. (1999). Decision support systems in action: integrated application in a multicriteria decision aid process. *European Journal of Operational Research*, 113(2):315–335.

Bana e Costa, C. A. and Vansnick, J.-C. (2008). A critical analysis of the eigenvalue method used to derive priorities in AHP. *European Journal of Operational Research*, 187(3):1422–1428.

Bansal, A. and Kumar, P. (2013). 3PL selection using hybrid model of AHP-PROMETHEE. *International Journal of Services and Operations Management*, 14(3):373–397.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barabási, A.-L. and Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288:60–69.

Barabási, A.-L. et al. (2009). Scale-free networks: A decade and beyond. *Science*, 325(5939):412.

Barnes, J. A. (1954). *Class and committees in a Norwegian island parish*. Plenum.

Barrat, A., Barthelemy, M., and Vespignani, A. (2008). *Dynamical processes on complex networks*. Cambridge University Press.

Barzilai, J., Cook, W. D., and Golany, B. (1987). Consistent weights for judgements matrices of the relative importance of alternatives. *Operations Research Letters*, 6(3):131–134.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the 2009 International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

Baumes, J., Goldberg, M., and Magdon-Ismail, M. (2005). Efficient identification of overlapping communities. In Kantor, P., Muresan, G., Roberts, F., Zeng, D. D., Wang, F.-Y., Chen, H., and Merkle, R. C., editors, *Intelligence and Security Informatics*, pages 27–36. Springer Berlin Heidelberg.

Beck, F., Burch, M., Diehl, S., and Weiskopf, D. (2014). The state of the art in visualizing dynamic graphs. In *Proceedings of the 2014 Eurographics Conference on Visualization – State of The Art Reports*. The Eurographics Association.

Behzadian, M., Kazemzadeh, R., Albadvi, A., and Aghdasi, M. (2010). PROMETHEE: A comprehensive literature review on methodologies and applications. *European Journal of Operational Research*, 200(1):198–215.

Belton, V. and Gear, T. (1983). On a short-coming of Saaty's method of analytic hierarchies. *Omega*, 11(3):228–230.

Belton, V. and Stewart, T. J. (2002). *Multiple criteria decision analysis: an integrated approach*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Bender-deMoll, S. and McFarland, D. A. (2006). The art and science of dynamic network visualization. *Journal of Social Structure*, 7(2):1–38.

Berger-Wolf, T. Y. and Saia, J. (2006). A framework for analysis of dynamic social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 523–528. ACM.

Berry, F. S., Brower, R. S., Choi, S. O., Goa, W. X., Jang, H., Kwon, M., and Word, J. (2004). Three traditions of network research: What the public management research agenda can learn from other research communities. *Public Administration Review*, 64(5):539–552.

Blenko, M. W., Mankins, M. C., and Rogers, P. (2010). The decision-driven organization. *Harvard Business Review*, 88(6):54–62.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175–308.

Bogdanovic, D., Nikolic, D., and Ilic, I. (2012). Mining method selection by integrated AHP and PROMETHEE method. *Anais da Academia Brasileira de Ciências*, 84(1):219–233.

Bonacich, P. (1987). Power and centrality: a family of measures. *The American Journal of Sociology*, 92(5):1170–1182.

Borgatti, S. P. and Foster, P. C. (2003). The network paradigm in organizational research: A review and typology. *Journal of Management*, 29(6):991–1013.

Borgatti, S. P. and Li, X. (2009). On social network analysis in a supply chain context. *Journal of Supply Chain Management*, 45(2):5–22.

Böttcher, M., Spott, M., Nauck, D., and Kruse, R. (2009). Mining changing customer segments in dynamic markets. *Expert Systems with Applications*, 36(1):155–164.

Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2007). On finding graph clusterings with maximum modularity. In Brandstädt, A., Kratsch, D., and Müller, H., editors, *Graph–heoretic Concepts in Computer Science*, volume 4769 of *Lecture Notes in Computer Science*, pages 121–132. Springer Berlin Heidelberg.

Brandes, U., Indlekofer, N., and Mader, M. (2012). Visualization methods for longitudinal social networks and stochastic actor-oriented modeling. *Social Networks*, 34(3):291–308.

Brandes, U., Kenis, P., and Raab, J. (2006). Explanation through network visualization. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(1):16–23.

Brandes, U. and Nick, B. (2011). Asymmetric relations in longitudinal social networks. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2283–2290.

Brans, J. P. (1982). L'ingéniérie de la décision: Elaboration d 'instruments d'aide à la décision. méthode PROMETHEE. In Nadeau, R. and Landry, M., editors, *Láide à la décision: nature, instruments et perspectives d'avenir*, pages 183–214. Presses de l'Université Laval, Québec, Canada.

Brans, J. P. and Mareschal, B. (1994). The PROMCALC and GAIA decision support system for MCDA. *Decision Support Systems*, 12(4-5):297–310.

Brans, J. P. and Vincke, P. (1985). A preference ranking organisation method (the PROMETHEE method for multiple criteria decision-making). *Management Science*, 31(6):647–656.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Brito, T. B., Silva, R. C. d. S., Botter, R. C., Pereira, N. N., and Medina, A. C. (2010). Discrete event simulation combined with multi-criteria decision analysis applied to steel plant logistics system planning. In *Proceedings of the 2010 Winter Simulation Conference*, pages 2126–2137. Winter Simulation Conference.

Bro, R. and Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics*, 17(1):16–33.

Brockenauer, R. and Cornelsen, S. (2001). Drawing clusters and hierarchies. In *Drawing graphs*, pages 193–227. Springer.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Computer Networks*, 33(1-6):309–320.

Bródka, P., Saganowski, S., and Kazienko, P. (2013). GED: the method for group evolution discovery in social networks. *Social Network Analysis and Mining*, 3(1):1–14.

Brynjolfsson, E., Hitt, L. M., and Kim, H. H. (2011). Strength in numbers: How does data-driven decision making affect firm performance? *Available at SSRN: http://ssrn.com/abstract=1819486*.

Burch, M., Schmidt, B., and Weiskopf, D. (2013). A matrix-based visualization for exploring dynamic compound digraphs. In *Proceedings of the 17th International Conference on Information Visualisation (IV)*, pages 66–73. IEEE.

Burke, L. A. and Miller, M. K. (1999). Taking the mystery out of intuitive decision making. *The Academy of Management Executive*, 13(4):91–99.

Burt, R. S. (1993). The social structure of competition. In Swedberg, R., editor, *Explorations in economic sociology*, pages 57–91. Russell Sage Foundation.

Camacho, J., Guimerà, R., and Amaral, L. A. N. (2002). Robust patterns in food web structure. *Physical Review Letters*, 88(22):228102.

Carley, K. M. (2003). *Dynamic network analysis*. Citeseer.

Carroll, J. D. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika*, 35(3):283–319.

Chakrabarti, D., Faloutsos, C., and McGlohon, M. (2010). Graph mining: Laws and generators. In Aggarwal, C. C. and Wang, H., editors, *Managing and Mining Graph Data*, pages 69–123. Springer.

Chakrabarti, D., Kumar, R., and Tomkins, A. (2006). Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 554–560. ACM.

Chang, Y.-H., Yeh, C.-H., and Chang, Y.-W. (2013). A new method selection approach for fuzzy group multicriteria decision making. *Applied Soft Computing*, 13(4):2179–2187.

Chen, C. and Morris, S. (2003). Visualizing evolving networks: Minimum spanning trees versus Pathfinder networks. In *Proceedings of the 2003 IEEE Symposium on Information Visualization*, pages 67–74. IEEE.

Chew, P. A., Bader, B. W., Kolda, T. G., and Abdelali, A. (2007). Cross-language information retrieval using PARAFAC2. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 143–152. ACM.

Chi, Y., Song, X., Zhou, D., Hino, K., and Tseng, B. L. (2009). On evolutionary spectral clustering. *ACM Transactions on Knowledge Discovery from Data*, 3(4):1–30.

Christley, R. M., Pinchbeck, G., Bowers, R., Clancy, D., French, N., Bennett, R., and Turner, J. (2005). Infection in social networks: using network analysis to identify high-risk individuals. *American Journal of Epidemiology*, 162(10):1024–1031.

Chung, K. H., Cho, K. K., Baek, T. H., and Park, J. C. (2001). An operation scheduling system for paint shop in the shipbuilding industry. *International Journal of Industrial Engineering: Theory, Applications and Practice*, 8(2):91–104.

Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6):066111.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Csardi, G. and Nepusz, T. (2006). The *igraph* software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9.

Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., and Joshi, A. (2008). Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, pages 668–677. ACM.

Datar, M., Gionis, A., Indyk, P., and Motwani, R. (2002). Maintaining stream statistics over sliding windows. *SIAM Journal on Computing*, 31(6):1794–1813.

Dağdeviren, M. (2008). Decision making in equipment selection: An integrated approach with AHP and PROMETHEE. *Journal of Intelligent Manufacturing*, 19(4):397–406.

de Blasio, B. F., Svensson, Å., and Liljeros, F. (2007). Preferential attachment in sexual networks. *Proceedings of the National Academy of Sciences*, 104(26):10762–10767.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Demirtas, E. A. and Üstün, O. (2008). An integrated multiobjective decision making process for supplier selection and order allocation. *Omega*, 36(1):76–90.

Derényi, I., Palla, G., and Vicsek, T. (2005). Clique percolation in random networks. *Physical Review Letters*, 94(16):160202.

Diestel, R. (2005). *Graph Theory*. Springer, 3rd edition.

Dodds, P. S., Muhamad, R., and Watts, D. J. (2003). An experimental study of search in global social networks. *Science*, 301(5634):827–829.

Domingos, P. (2005). Mining social networks for viral marketing. *IEEE Intelligent Systems*, 20(1):80–82.

Dooley, A. E., Smeaton, D. C., Sheath, G. W., and Ledgard, S. F. (2009). Application of multiple criteria decision analysis in the New Zealand agricultural industry. *Journal of Multi-Criteria Decision Analysis*, 16(1-2):39–53.

Drucker, P. F., Drucker, P. F., and Drucker, P. F. (1967). *The effective executive*, volume 967. Heinemann London.

Dunlavy, D. M., Kolda, T. G., and Acar, E. (2011). Temporal link prediction using matrix and tensor factorizations. *ACM Transactions on Knowledge Discovery from Data*, 5(2):10.

Dyer, J. S. (1990). Remarks on the Analytic Hierarchy Process. *Management Science*, 36(3):249–258.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds and markets: reasoning about a highly connected world*. Cambridge University Press, Cambridge.

Eckenrode, R. T. (1965). Weighting multiple criteria. *Management Science*, 12(3):180–192.

Eisenberg, E. and Levanon, E. Y. (2003). Preferential attachment in the protein network evolution. *Physical Review Letters*, 91(13):138701.

Elliot, T. (2012). The year analytics means business. `http://www.smartdatacollective.com/timoelliott/45868/2012-year-analytics-means-business`. Posted: 10-02-2012.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1960):17–61.

Evans, J. R. and Lindner, C. H. (2012). Business analytics: the next frontier for decision sciences. *Decision Line*, 43(2):4–6.

Evans, T. and Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1):016105.

Falkowski, T., Bartelheimer, J., and Spiliopoulou, M. (2006). Mining and visualizing the evolution of subgroups in social networks. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 52–58. IEEE.

Ferlez, J., Faloutsos, C., Leskovec, J., Mladenic, D., and Grobelnik, M. (2008). Monitoring network evolution using MDL. In *Proceedings of the IEEE 24th International Conference on Data Engineering*, pages 1328–1330. IEEE.

Forman, E. and Peniwati, K. (1998). Aggregating individual judgments and priorities with the Analytic Hierarchy Process. *European Journal of Operational Research*, 108(1):165–169.

Forman, E. H. and Gass, S. I. (2001). The Analytic Hierarchy Process - an exposition. *Operations Research*, 49(4):469–486.

Fortunato, S. (2010). Community detection in graphs. *Physics Report*, 486(3-5):75–174.

Fortunato, S. and Castellano, C. (2012). Community structure in graphs. In Meyers, R. A., editor, *Computational Complexity*, pages 490–512. Springer.

Freeman, L. C. (1979). Centrality in social networks: conceptual clarification. *Social Networks*, 1(3):215–239.

Fruchterman, T. M. J. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164.

Fu, W., Song, L., and Xing, E. P. (2009). Dynamic mixed membership blockmodel for evolving networks. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 329–336. IMLS.

Gabaix, X. (2009). Power laws in economics and finance. *Annual Review of Economics*, 1(1):255–294.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144.

Gartner (2015). Hype cycle for emerging technologies 2013. `http://www.gartner.com/newsroom/id/2575515`. Accessed: 20-08-2015.

Geffen, C. A. and Rothenberg, S. (2000). Suppliers and environmental innovation: the automotive paint process. *International Journal of Operations & Production Management*, 20(2):166–186.

Gehrke, J., Korn, F., and Srivastava, D. (2001). On computing correlated aggregates over continual data streams. *ACM SIGMOD Record*, 30(2):13–24.

Gervásio, H. and Simões da Silva, L. (2012). A probabilistic decision-making approach for the sustainable assessment of infrastructures. *Expert Systems with Applications*, 39(8):7121–7131.

Ghoniem, M., Fekete, J. D., and Castagliola, P. (2005). On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization*, 4(2):114–135.

Gibson, H., Faith, J., and Vickers, P. (2013). A survey of two-dimensional graph layout techniques for information visualisation. *Information Visualization*, 12(3-4):324–357.

Giordani, P., Kiers, H., and Del Ferraro, M. (2012). Three-way component analysis using the R package *threeway*.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826.

Gleiser, P. M. and Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems*, 6(04):565–573.

Goodwin, P. and Wright, G. (2004). *Decision analysis for management judgment*. John Wiley & Sons, Chichester, UK, 3rd edition.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

Granovetter, M. S. (1974). *Getting a job: a study of contacts and careers*. Harvard University Press, Cambridge, MA.

Greene, D., Doyle, D., and Cunningham, P. (2010). Tracking the evolution of communities in dynamic social networks. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 176–183. IEEE.

Gregory, S. (2007). An algorithm to find overlapping community structure in networks. In Kok, J. N., Koronacki, J., Mantaras, R. L., Matwin, S., Mladenič, D., and Skowron, A., editors, *Knowledge discovery in databases: PKDD 2007*, pages 91–102. Springer.

Gregory, S. (2010). Finding overlapping communities in networks by Label Propagation. *New Journal of Physics*, 12(10):103018.

Guimera, R. and Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900.

Gupta, M., Aggarwal, C. C., Han, J., and Sun, Y. (2011). Evolutionary clustering and analysis of bibliographic networks. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 63–70. IEEE.

Gupta, S., Anderson, R. M., and May, R. M. (1989). Networks of sexual contacts: implications for the pattern of spread of HIV. *AIDS*, 3(12):807–818.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: models and conditions for an explanatory multi-modal factor analysis. Technical report, UCLA - University of California, Los Angeles. Working Paper in Phonetics.

Harshman, R. A. (1972). PARAFAC2: Mathematical and technical notes. *UCLA Working Papers in Phonetics*, 22(3044):122215.

Hastie, R. and Dawes, R. (2001). *Rational Choice in an Uncertain World: The Psychology of Judgement and Decision Making*. SAGE Publications.

Henry, N. and Fekete, J.-D. (2006). Matrixexplorer: a dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):677–684.

Herva, M. and Roca, E. (2013). Ranking municipal solid waste treatment alternatives based on ecological footprint and multi-criteria analysis. *Ecological Indicators*, 25(2013):77–84.

Hill, S., Provost, F., and Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276.

Hobbs, B. F. and Meier, P. (2012). *Energy decisions and the environment: a guide to the use of multicriteria methods*, volume 28. Springer Science & Business Media.

Holland, P. W. and Leinhardt, S. (1977). A dynamic model for social networks. *Journal of Mathematical Sociology*, 5(1):5–20.

Holsapple, C., Lee-Post, A., and Pakath, R. (2014). A unified foundation for business analytics. *Decision Support Systems*, 64:130–141.

Hopcroft, J., Khan, O., Kulis, B., and Selman, B. (2004). Tracking evolving communities in large linked networks. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5249–5253.

Huang, Z. and Lin, D. K. (2009). The time-series link prediction problem with applications in communication surveillance. *INFORMS Journal on Computing*, 21(2):286–303.

Huberman, B. A. and Adamic, L. A. (1999). Internet: growth dynamics of the World-Wide Web. *Nature*, 401(6749):131–131.

Ilangkumaran, M., Sasirekha, V., Anojkumar, L., Sakthivel, G., Raja, M. B., Raj, T. R. S., Siddhartha, C., Nizamuddin, P., and Kumar, S. P. (2013). Optimization of wastewater treatment technology selection using hybrid MCDM. *Management of Environmental Quality: An International Journal*, 24(5):619–641.

INFORMS (2015). What is analytics? `https://www.informs.org/About-INFORMS/What-is-Analytics`. Accessed: 19-08-2015.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666.

Jeong, H., Néda, Z., and Barabási, A.-L. (2003). Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.

Kabak, M., Burmaoğlu, S., and Kazançoğlu, Y. (2012). A fuzzy hybrid MCDM approach for professional selection. *Expert Systems with Applications*, 39(3):3516–3525.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15.

Kara, S. S. (2011). Evaluation of outsourcing companies of waste electrical and electronic equipment recycling. *International Journal of Environmental Science & Technology*, 8(2):291–304.

Kawadia, V. and Sreenivasan, S. (2012). Sequential detection of temporal communities by estrangement confinement. *Scientific Reports*, 2(794):1–10.

Keeney, R. L. (1982). Decision analysis: an overview. *Operations Research*, 30(5):803–838.

Kernighan, B. W. and Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49(2):291–307.

Keune, H., Springael, J., and Keyser, W. D. (2013). Negotiated complexity: framing multi-criteria decision support in environmental health practice. *American Journal of Operations Research*, 3(1):153–166.

Khayamian, T., Sajjadi, S., Mirmahdieh, S., Mardihallaj, A., and Hashemian, Z. (2012). Simultaneous analysis of bifenthrin and tetramethrin using corona discharge ion mobility spectrometry and tucker 3 model. *Chemometrics and Intelligent Laboratory Systems*, 118:88–96.

Kiers, H. A. (1998). Joint orthomax rotation of the core and component matrices resulting from three-mode principal components analysis. *Journal of Classification*, 15(2):245–263.

Kiers, H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14(3):105–122.

Kiers, H. A. L. and Van Mechelen, I. (2001). Three-way component analysis: principles and illustrative application. *Psychological Methods*, 6(1):84–110.

Kim, M.-S. and Han, J. (2009). A particle-and-density based evolutionary clustering method for dynamic networks. *Proceedings of the VLDB Endowment*, 2(1):622–633.

Kiron, D., Shockley, R., Kruschwitz, N., Finch, G., and Haydock, M. (2011). Analytics: The widening divide. *MIT Sloan Management Review*, 53(3):1–22.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The web as a graph: measurements, models, and methods. In Asano, T., Imai, H., Lee, D., Nakano, S.-i., and Tokuyama, T., editors, *Computing and Combinatorics*, volume 1627 of *Lecture Notes in Computer Science*, pages 1–17. Springer.

Kolda, T. G. and Bade, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Korte, C. and Milgram, S. (1970). Acquaintance networks between racial groups: Application of the small world method. *Journal of Personality and Social Psychology*, 15(2):101.

Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.

Kroonenberg, P. M. (1983). *Three-mode principal component analysis: theory and applications*. DSWO Press.

Kroonenberg, P. M. (2008). *Applied multiway data analysis*, volume 702. John Wiley & Sons.

Kroonenberg, P. M. and De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 45(1):69–97.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review E*, 80(5):056117.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., and Kruschwitz, N. (2013). Big data, analytics and the path from insights to value. *MIT Sloan Management Review*, 21.

Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics & Data Analysis*, 18(1):97–119.

Lee, B. (2012). Marketing is dead. *Harvard Business Review*, pages 1–3.

Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., and Henry, N. (2006). Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, pages 1–5. ACM.

Leicht, E. A. and Newman, M. E. (2008). Community structure in directed networks. *Physical Review Letters*, 100(11):118703.

Leichtling, C. (2002). How color affects marketing. *Touro College Accounting and Business Society Journal*, 2:22–31.

Leskovec, J., Adamic, L., and Huberman, B. (2007a). The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1):1–39.

Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008a). Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 462–470. ACM.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 177–187. ACM.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007b). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1):2.

Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2008b). Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th International Conference on World Wide Web*, pages 695–704. ACM.

Levin, J. (1965). Three-mode factor analysis. *Psychological Bulletin*, 64(6):442–452.

Levine, J. H. (1972). The sphere of influence. *American Sociological Review*, 37(1):14–27.

Lewin, K., Heider, F. T., and Heider, G. M. (1936). *Principles of topological psychology*. McGraw-Hill, New York, NY, US, 1st edition.

Li, J., Blumenfeld, D. E., and Marin, S. P. (2007). Manufacturing system design to improve quality buy rate: an automotive paint shop application study. *IEEE Transactions on Automation Science and Engineering*, 4(1):75–79.

Lin, Y.-R., Chi, Y., Zhu, S., Sundaram, H., and Tseng, B. L. (2009). Analyzing communities and their evolutions in dynamic social networks. *ACM Transactions on Knowledge Discovery from Data*, 3(2):8:1–8:31.

Lorrain, F. and White, H. C. (1971). Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, 1(1):49–80.

Lu, J., Zhang, G., Ruan, D., and Wu, F. (2007). *Multi-objective group decision making-methods, software and applications with fuzzy set technology*, volume 6 of *Series in Electrical and Computer Engineering*. Imperial College Press, London, UK.

Luce, R. D. and Raiffa, H. (1957). *Games and decisions: Introduction and critical surveys*. Wiley New York.

Macharis, C., Springael, J., Brucker, K. D., and Verbeke, A. (2004). PROMETHEE and AHP: The design of operational synergies in multicriteria analysis. strengthening PROMETHEE with ideas of AHP. *European Journal of Operational Research*, 153(2):307–317.

Mashima, D., Kobourov, S. G., and Hu, Y. (2012). Visualizing dynamic data with maps. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1424–1437.

McGlohon, M., Akoglu, L., and Faloutsos, C. (2011). Statistical properties of social networks. In Aggarwal, C. C., editor, *Social Network Data Analytics*, pages 17–42. Springer.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444.

Mei, Q. and Zhai, C. (2005). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 198–207. ACM.

Meunier, F. and Neveu, B. (2012). Computing solutions of the paintshop-necklace problem. *Computers & Operations Research*, 39(11):2666–2678.

Miguéis, V. L., Camanho, A. S., and Falcão e Cunha, J. (2012). Customer data mining for lifestyle segmentation. *Expert Systems with Applications*, 39(10):9359–9366.

Milgram, S. (1967). The small world problem. *Psychology Today*, 1(1):61–67.

Miller, C. C. and Ireland, R. D. (2005). Intuition in strategic decision making: friend or foe in the fast-paced 21st century? *The Academy of Management Executive*, 19(1):19–30.

Moody, J., McFarland, D., and Bender-deMoll, S. (2005). Dynamic network visualization. *American Journal of Sociology*, 110(4):1206–1241.

Moreno, J. L. (1934). *Who Shall Survive?* Beacon House, New York, NY, US, 1st edition.

Mørup, M. and Hansen, L. K. (2009). Automatic relevance determination for multi-way models. *Journal of Chemometrics*, 23(7-8):352–363.

Murphy, C. K. (1993). Limits on the Analytic Hierarchy Process from its consistency index. *European Journal of Operational Research*, 65(1):138–139.

Nanavati, A. A., Singh, R., Chakraborty, D., Dasgupta, K., Mukherjea, S., Das, G., Gurumurthy, S., and Joshi, A. (2008). Analyzing the structure and evolution of massive telecom graphs. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):703–718.

Nasiri, H., Boloorani, A. D., Sabokbar, H. A. F., Jafari, H. R., Hamzeh, M., and Rafii, Y. (2013). Determining the most suitable areas for artificial groundwater recharge via an integrated PROMETHEE II-AHP method in GIS environment (case study: Garabaygan Basin, Iran). *Environmental Monitoring and Assessment*, 185(1):707–718.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351.

Newman, M. E. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409.

Newman, M. E. J. (2003a). Mixing patterns in networks. *Physical Review E*, 67(2):026126.

Newman, M. E. J. (2003b). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113.

Nicosia, V., Mangioni, G., Carchiolo, V., and Malgeri, M. (2009). Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(03):P03024.

Ogawa, M. and Ma, K.-L. (2010). Software evolution storylines. In *Proceedings of the 5th International Symposium on Software Visualization*, pages 35–42. ACM.

Oliveira, M. and Gama, J. (2012). A framework to monitor clusters evolution applied to economy and finance problems. *Intelligent Data Analysis*, 16(1):93–111.

Oliveira, M., Guerreiro, A., and Gama, J. (2014). Dynamic communities in evolving customer networks: an analysis using landmark and sliding windows. *Social Network Analysis and Mining*, 4(1):1–19.

Opsahl, T., Agneessens, F., and Skvoretz, J. (2010). Node centrality in weighted networks: generalizing degree and shortest paths. *Social Networks*, 32(3):245–251.

Opsahl, T. and Panzarasa, P. (2009). Clustering in weighted networks. *Social Networks*, 31(2):155–163.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

Palla, G., Barabási, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664–667.

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Pattillo, J., Youssef, N., and Butenko, S. (2013). On clique relaxation models in network analysis. *European Journal of Operational Research*, 226(1):9–18.

Patton, J. R. (2003). Intuition in decisions. *Management Decision*, 41(10):989–996.

Perny, P. (1998). Multicriteria filtering methods based on concordance and non-discordance principles. *Annals of Operations Research*, 80(1998):137–165.

Petersen, A. M., Jung, W.-S., Yang, J.-S., and Stanley, H. E. (2011). Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences*, 108(1):18–23.

Philpott, S. (2010). Advanced analytics: Unlocking the power of insight. *IBM, April*.

Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In *Proceedings of the 20th International Symposium on Computer and Information Sciences*, volume 3733, pages 284–293. Springer Berlin Heidelberg.

Powell, W. W., Koput, K. W., and Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly*, 41(1):116–145.

Praseeda, C. and Shivakumar, B. (2014). A review of trends and technologies in business analytics. *International Journal of Advanced Research in Computer Science*, 5(8).

Price, D. D. S. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663.

Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106.

Rapoport, A. (1953). Spread of information through a population with socio-structural bias I: assumption of transitivity. *The Bulletin of Mathematical Biophysics*, 15(4):523–533.

Reagans, R., Zuckerman, E., and McEvily, B. (2004). How to make the team: Social networks vs. demography as criteria for designing effective teams. *Administrative Science Quarterly*, 49(1):101–133.

Reda, K., Tantipathananandh, C., Johnson, A., Leigh, J., and Berger-Wolf, T. (2011). Visualizing the evolution of community structures in dynamic social networks. *Computer Graphics Forum*, 30(3):1061–1070.

Richardson, M. and Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 61–70. ACM.

Ritter, T. (1999). The networking company: antecedents for coping with relationships and networks effectively. *Industrial Marketing Management*, 28(5):467–479.

Rombach, M. P., Porter, M. A., Fowler, J. H., and Mucha, P. J. (2014). Core-periphery structure in networks. *SIAM Journal on Applied Mathematics*, 74(1):167–190.

Roodposhti, M. S., Rahimi, S., and Beglou, M. J. (2014). PROMETHEE II and fuzzy AHP: an enhanced GIS-based landslide susceptibility mapping. *Natural Hazards*, 73(1):77–95.

Roy, B. (1968). Classement et choix en présence de points de vue multiples. *Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle*, 2(1):57–75.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3):234–281.

Saaty, T. L. (1986). Axiomatic foundation of the Analytic Hierarchy Process. *Management Science*, 32(7):841–855.

Saaty, T. L. (1990). How to make a decision: the Analytic Hierarchy Process. *European Journal of Operational Research*, 48(1):9–26.

Saganowski, S., Bródka, P., and Kazienko, P. (2012). Influence of the dynamic social network timeframe type and size on the group evolution discovery. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*, pages 679–683. IEEE.

Sallaberry, A., Muelder, C., and Ma, K.-L. (2013). Clustering, visualizing, and navigating for large dynamic graphs. In Didimo, W. and Patrignani, M., editors, *Graph Drawing*, pages 487–498. Springer Berlin Heidelberg.

Samtaney, R., Silver, D., Zabusky, N., and Cao, J. (1994). Visualizing features and tracking their evolution. *Computer*, 27(7):20–27.

Samuelson, P. A. (1938). The empirical implications of utility analysis. *Econometrica*, 6(4):344–356.

Sarkar, P., Chakrabarti, D., and Jordan, M. (2012). Nonparametric link prediction in dynamic networks. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1687–1694.

Sarkar, P. and Moore, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 7(2):31–40.

Schoner, B. S. and Wedley, B. W. (1989). Ambiguous criteria weights in AHP: Consequences and solutions. *Decision Sciences*, 20(3):462–475.

Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P., Mukherjee, G., and Manna, S. (2003). Small-world properties of the Indian railway network. *Physical Review E*, 67(3):036106.

Seydel, J. (2006). Data envelopment analysis for decision support. *Industrial Management & Data Systems*, 106(1):81–95.

Shetty, J. and Adibi, J. (2004). Email dataset database schema and brief statistical report. Technical report, Federal Energy Regulatory Commission.

Shrum, W., Cheek Jr, N. H., and MacD, S. (1988). Friendship in school: Gender and racial homophily. *Sociology of Education*, 61:227–239.

Shyur, H.-J. and Shih, H.-S. (2006). A hybrid MCDM model for strategic vendor selection. *Mathematical and Computer Modelling*, 44(7-8):749–761.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, 1(1):161–176.

Simon, H. A. (1987). Making management decisions: The role of intuition and emotion. *The Academy of Management Executive*, 1(1):57–64.

Simon, H. A., Dantzig, G. B., Hogarth, R., Plott, C. R., Raiffa, H., Schelling, T. C., Shepsle, K. A., Thaler, R., Tversky, A., and Winter, S. (1987). Decision making and problem solving. *Interfaces*, 17(5):11–31.

Skillicorn, D. (2007). *Understanding complex datasets: data mining with matrix decompositions*. Chapman and Hall/CRC.

Skillicorn, D., Zheng, Q., and Morselli, C. (2014). Modeling dynamic social networks using spectral embedding. *Social Network Analysis and Mining*, 4(1):1–14.

Smilde, A. K. (1992). Three-way analyses problems and prospects. *Chemometrics and Intelligent Laboratory Systems*, 15(2):143–157.

Snijders, T. A., Van de Bunt, G. G., and Steglich, C. E. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60.

Spiliopoulou, M. (2011). Evolution in social networks: A survey. In Aggarwal, C. C., editor, *Social Network Data Analytics*, pages 149–175. Springer.

Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., and Schult, R. (2006). MONIC: modeling and monitoring cluster transitions. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 706–711. ACM.

Stephen, A. T. and Toubia, O. (2009). Explaining the power-law degree distribution in a social commerce network. *Social Networks*, 31(4):262–270.

Stewart, T. J. (1992). A critical survey on the status of multiple criteria decision making theory and practice. *Omega*, 20(5):569–586.

Sun, J., Faloutsos, C., Papadimitriou, S., and Yu, P. S. (2007). Graphscope: parameter-free mining of large time-evolving graphs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 687–696. ACM.

Sun, J., Papadimitriou, S., Lin, C., Cao, N., Liu, S., and Qian, W. (2009). Multivis: content-based social network exploration through multi-way visual analysis. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 1063–1074. SIAM.

Sun, J., Tao, D., Papadimitriou, S., Yu, P. S., and Faloutsos, C. (2008). Incremental tensor analysis: theory and applications. *ACM Transactions on Knowledge Discovery from Data*, 2(3):11:1–11:37.

Taha, Z. and Rostam, S. (2012). A hybrid fuzzy AHP-PROMETHEE decision support system for machine tool selection in flexible manufacturing cell. *Journal of Intelligent Manufacturing*, 23(6):2137–2149.

Takaffoli, M., Rabbany, R., and Zaïane, O. R. (2013). Incremental local community identification in dynamic social networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94. ACM.

Takaffoli, M., Sangi, F., Fagnan, J., and Zaïane, O. R. (2011). Community evolution mining in dynamic social networks. *Procedia - Social and Behavioral Sciences*, 22(2011):49–58.

Tanahashi, Y. and Ma, K.-L. (2012). Design considerations for optimizing storyline visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2679–2688.

Tantipathananandh, C., Berger-Wolf, T., and Kempe, D. (2007). A framework for community identification in dynamic social networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 717–726. ACM.

Team, R. D. C. (2008). R: A language and environment for statistical computing. Technical report, R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0.

Thagard, P. (2001). How to make decisions: coherence, emotion, and practical inference. In Millgram, E., editor, *Varieties of practical inference*, pages 355–371. MIT Press: Cambridge, MA.

Thelwall, M. (2006). Interpreting social science link analysis research: a theoretical framework. *Journal of the American Society for Information Science and Technology*, 57(1):60–68.

Timmerman, M. E. and Kiers, H. A. L. (2000). Three mode principal component analysis: choosing the numbers of components and sensitivity to local optima. *British Journal of Mathematical and Statistical Psychology*, 53(1):1–16.

Toyoda, M. and Kitsuregawa, M. (2003). Extracting evolution of web communities from a series of web archives. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia*, pages 28–37. ACM.

Travers, J. and Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 1(1):425–443.

Tucker, L. (1963). Implications of factor analysis of three-way matrices for measurement of change. In Harris, C. W., editor, *Problems in Measuring Change*, pages 122–137. John Wiley & Sons.

Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Tufte, E. R. and Graves-Morris, P. (1983). *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT.

Turcksin, L., Bernardini, A., and Macharis, C. (2011). A combined AHP-PROMETHEE approach for selecting the most appropriate policy scenario to stimulate a clean vehicle fleet. *Procedia - Social and Behavioral Sciences*, 20(2011):954–965.

Turskis, Z. and Zavadskas, E. K. (2011). Multiple criteria decision making (MCDM) methods in economics: an overview. *Technological and Economic Development of Economy*, 2(2011):397–427.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Tylenda, T., Angelova, R., and Bedathur, S. (2009). Towards time-aware link prediction in evolving social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, pages 1–9. ACM.

Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*.

Ulgen, U. and Gunal, A. (1998). Simulation in the automobile industry. In Banks, J., editor, *Handbook of Simulation*, pages 547–570. John Wiley & Sons.

Valente, T. W. and Foreman, R. K. (1998). Integration and radiality: measuring the extent of an individual's connectedness and reachability in a network. *Social Networks*, 20(1):89–105.

Van De Bunt, G. G., Van Duijn, M. A. J., and Snijders, T. A. B. (1999). Friendship networks through time: an actor-oriented dynamic statistical network model. *Computational & Mathematical Organization Theory*, 5(2):167–192.

van den Elzen, S., Holten, D., Blaas, J., and van Wijk, J. J. (2014). Dynamic network visualization with extended massive sequence views. *IEEE Transactions on Visualization and Computer Graphics*, 20(8):1087–1099.

Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H., and Chklovskii, D. B. (2011). Structural properties of the Caenorhabditis elegans neuronal network. *PLoS Computational Biology*, 7(2):e1001066.

Vehlow, C., Burch, M., Schmauder, H., and Weiskopf, D. (2013). Radial layered matrix visualization of dynamic graphs. In *Proceedings of the 17th International Conference on Information Visualisation (IV)*, pages 51–58. IEEE.

Venkatesan, S. P. and Kumanan, S. (2012). Supply chain risk prioritisation using a hybrid AHP and PROMETHEE approach. *International Journal of Services and Operations Management*, 13(1):19–41.

Von Neumann, J. and Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Von Winterfeldt, D. and Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge University Press, Cambridge, New York, USA, 1st edition.

Wang, C.-D., Lai, J.-H., and Yu, P. S. (2014). NEIWalk: Community discovery in dynamic content-based networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(7):1734–1748.

Wang, J., Li, J., and Huang, N. (2009). Optimal scheduling to achieve energy reduction in automotive paint shops. In *ASME 2009 International Manufacturing Science and Engineering Conference*, pages 161–167. American Society of Mechanical Engineers.

Wang, J.-J. and Yang, D.-L. (2007). Using a hybrid multi-criteria decision aid method for information systems outsourcing. *Computers & Operations Research*, 34(12):3691–3700.

Wang, L., Rege, M., Dong, M., and Ding, Y. (2012). Low-rank kernel matrix factorization for large-scale evolutionary clustering. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1036–1050.

Wang, Y.-M., Liu, J., and Elhag, T. M. (2008). An integrated AHP-DEA methodology for bridge risk assessment. *Computers & Industrial Engineering*, 54(3):513–525.

Wasserman, S. (1980). Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, 75(370):280–294.

Wasserman, S. and Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge University Press, Cambridge.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440–442.

Wei, C.-P. and Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications*, 23(2):103–112.

White, D. R. and Reitz, K. P. (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5(2):193–234.

White, H. C. (1963). *An anatomy of kinship: Mathematical models for structures of cumulated roles*. Prentice-Hall Englewood Cliffs, NJ.

Xie, J., Kelley, S., and Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 45(4):43.

Xu, J. and Chen, H. (2005). Criminal network analysis and visualization. *Communications of the ACM*, 48(6):100–107.

Yang, T., Chi, Y., Zhu, S., Gong, Y., and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks - a Bayesian approach. *Machine Learning*, 82(2):157–189.

Yang, Y.-P. O., Shieh, H. M., Leu, J. D., and Tzeng, G. H. (2008). A novel hybrid MCDM model combined with DEMATEL and ANP with applications. *International Journal of Operations Research*, 5(3):160–168.

Yi, J. S., Elmqvist, N., and Lee, S. (2010). Timematrix: analyzing temporal social networks using interactive matrix-based visualizations. *International Journal of Human-Computer Interaction*, 26(11-12):1031–1051.

Zhou, T., Ren, J., Medo, M., and Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *Physical Review E*, 76(4):046115.

Zionts, S. (1979). MCDM: If not a roman numeral, then what? *Interfaces*, 9(4):94–101.